# NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

# OFFICE OF OCEAN EXPLORATION

# Data Management Strategy
# for the Ocean Exploration Program

## April 30, 2002

MTS MitreTek Systems Innovative Technology in the Public Interest™
3150 Fairview Park Drive South , Falls Church, Virginia 22042-4519
703.610.2000; www.mitretek.org

# NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

# OFFICE OF OCEAN EXPLORATION


# Data Management Strategy for the Ocean Exploration Program

**Contract Number: 50-SPNA-9-00009**
**Task Order: 56-SPNA-9-90020**

**Contributing Authors: Gary M. Mineart, Fred C. Klein, Miro Medek, Michael J. Ciarametaro, Bradford T. Ulery, Stephen M. Holt, Albert E. Fletcher, Amy E. Sheridan**

## April 30, 2002

# TABLE OF CONTENTS

# LIST OF FIGURES



# LIST OF TABLES

# EXECUTIVE SUMMARY

This document provides a strategy for the management of data obtained through the NOAA ocean exploration program. When implemented, it will support the Office of Ocean Exploration (OE) by providing a viable and efficient means for managing, using, and providing access to data collected during OE program-sponsored activities. The OE program and all ocean exploration stakeholders need an ability to systematically manage and exploit the data generated during these activities in order to assess, catalog, document, and preserve their findings.

Principles that govern the management of ocean exploration data result from the OE program vision articulated in the program's strategic framework and based on the October 2000 Report of the President's Panel on Ocean Exploration. These principles place requirements on OE to manage the collection and distribution of large volumes of ocean exploration data, and lead to the following business driver:

> *Collect and Distribute Ocean Exploration Information*. Collect data and information from ocean exploration activities and share this information in such a way that is available to all stakeholders, including the general public.

Data management goals and objectives contained within this strategy support the business driver and set targets related to data collection, storage, access, and archival. Guiding principles also establish the need for a distributed approach to ocean exploration data management to satisfy the requirements driven by the variety and complexity of oceanographic data types and the diversity of stakeholders.

The environment in which OE must develop a data management capability is one of rapid change, conflicting guidance on proprietary data, and an explosion of data complexity and quantity. The growing challenges of this data environment are being addressed by an abundance of information systems solutions bounded by policies of national need, security, and proprietary rights. Assumptions, constraints, and enabling technologies within this environment define the range of data management solutions that can be applied. An important component of the management environment is the application of

metadata standards and practices by OE program participants. The data management process within this environment is applicable to a diverse mix of oceanographic data types and volumes of data approaching one terabyte (TB) for the current year exploration campaign and nearly 90 TB over the next 10 years. Functional areas of the data management process are illustrated in an ocean exploration data flow model that tracks the data lifecycle from collection to archival. The data flow model accommodates a flexible mix of data types, catalogs, and both centralized and distributed storage. Policies that apply to this data management environment provide for access to these data by a broad cross-section of potential users, guided by the following principles:

- Ocean exploration is an investment in the public interest
- Discoveries from ocean exploration rely on full and open access to data
- A market model for access to ocean exploration data is unsuitable for research, education, and outreach
- The interests of ocean exploration data owners must be balanced with society's need for open exchange of information

OE must coordinate with a diverse community of partners in its implementation of this data management strategy to maximize effectiveness and return on investment.

Data management architecture alternatives are examined within the context of the principles and environment. The functional components of the various alternatives are examined for the data flow model categories of collection, processing, storage, access, and archival. Three viable alternatives result from this examination:

- *Alternative 1:* OE central repository and catalog
- *Alternative 2:* OE distributed repository with centralized catalog
- *Alternative 3:* OE central repository and catalog with replication of data at host sites and discipline-specific research organizations

A general assessment of these alternatives and an examination of risk factors result in this strategy's high-level design recommendation based on Alternative 2. A concept of operations describes the multiple phases of the functional process, roles and responsibilities, operational procedures, and policies that support this alternative. The set of actions by OE to implement this strategy include the generation and distribution of data management policy guidance, communication with national-level stakeholders,

designation of cognizant staff, and establishment of partnerships with various components of the National Environmental Satellite, Data, and Information Service (NESDIS) to guide the development of catalog, repository, and archive systems capabilities.

# DATA MANAGEMENT STRATEGY FOR THE OCEAN EXPLORATION PROGRAM

This data management strategy has been prepared for the National Oceanic and Atmospheric Administration (NOAA) Office of Ocean Exploration (OE) by Mitretek Systems, Inc. Mitretek is a nonprofit corporation chartered to work in the public interest and is under a directed award contract with NOAA to provide objective, conflict-free advice, especially regarding information technology (IT) investment decision-making, program management, and budget and strategy formulation. This document is provided in accordance with and fulfillment of the Deliverable 4 requirement in Task 20 of the Mitretek contract with NOAA.

## 1 NEED FOR AN OE DATA MANAGEMENT SYSTEM

This section provides an overview of the need and justification for establishing a system—the infrastructure, supporting software, policy, and procedures—for the management of ocean exploration data. These data will be collected under the principal auspices of the Office of Ocean Exploration (OE) within the National Oceanic and Atmospheric Administration (NOAA). The strategy for managing ocean exploration data contained in this document is also applicable to other agencies involved in ocean exploration activities.

### 1.1 Introduction

NOAA established OE in 2001 to lead a revitalized national strategy of ocean exploration through implementation of a dedicated program for ocean exploration. The critical importance of ocean exploration to our understanding of the Earth is being recognized with increasing frequency and has received national attention through efforts such as the President's Panel on Ocean Exploration (whose guiding document, published in 2000, is hereafter referred to as the *Frontier Report*)[1] and testimony before Congress by NOAA leadership and other distinguished advocates. The *Frontier Report* defines exploration as "discovery through disciplined diverse observations and recording of the findings."[2] The U. S. Navy, a partner in the President's Panel process, further refined this definition into a widely accepted benchmark: "[Ocean exploration] is systematic examination for the purposes of discovery; cataloging and documenting what one finds; boldly going where no one has

gone before; providing an initial knowledge base for hypothesis-based science and for exploitation[3].” This definition recognizes that true exploration is planned, programmed, and executed for its own sake (not achieved opportunistically). It requires a known starting point, may have significant programmatic risk, has valuable practical applications, and emphasizes the recording of results to advance the known knowledge base across a broad user community. Clearly, those involved in exploration cannot embark on processes of systematic examination, cataloging, and documentation of their findings without the ability to systematically manage the data generated by these processes.

In support of the broad strategy contained in the *Frontier Report* and in its statutory role as the nation’s agency for ocean stewardship, NOAA initiated the OE program to coordinate involvement in exploration activities across the five NOAA line offices and other national stakeholders. As a principal component of this national strategy, the OE program seeks to bring the best of the nation’s scientists to the leading edges of ocean science and technology. In so doing, OE positions them to discover more about life and processes within the oceans and learn more about maritime cultural resources and heritage, thereby reaping the benefits of the ocean’s biological and mineral resources.

Every ocean expedition has the potential to discover important information about the origins of life on earth or about new living or non-living resources that may benefit humanity. Recent progress in technology is enabling new initiatives. Ocean exploration assets and capabilities may one day rival those of space exploration, with potential for enormous economic, archeological, health, and quality of life benefits.[4] New discoveries and the acquisition and dissemination of knowledge resulting from ocean exploration promise to enhance the regulatory effectiveness of government and industry. Better decision-making ability from an informed perspective will improve conservation, utilization, and management of ocean resources. Education and public outreach to increase interest and knowledge of the oceans through the excitement of exploration and discovery will lead to greater public awareness of ocean issues and a renewed appreciation of America’s maritime heritage. A widely supported national emphasis on oceanographic research, stewardship of ocean resources, and sensible commercial use will be achieved as the public’s knowledge of and

interest in the oceans increases over time. A rational and effective system of data collection, management, and access is essential if this vision is to become a reality.

New oceanographic sensors can collect more data in one hour then the HMS Challenger expedition of 1872-1876 collected in one year[5]. Data preservation and management are critical to the success of exploration, both for sharing discoveries made during the course of exploration activities and for facilitating further analysis and follow-on research. This subsequent analysis of exploration data by subject matter experts across a broad spectrum of disciplines will lead to additional discoveries and will help satisfy the goals and objectives of the research community. Ocean exploration data must be accessible to a large interdisciplinary community and archived for posterity to provide "a legacy of new knowledge that can be used by those not yet born to answer the questions not yet posed at the time of the exploration."[6] NOAA, the national ocean exploration community, and federal regulators must be prepared to embrace new principles for the management of exploration data.

Applying these principles to a data management strategy will foster a robust public dissemination and collaborative spirit to share and exploit exploration data.

## 1.2 Data Management Strategy Purpose

This document provides a strategy for NOAA management of ocean exploration data obtained through the OE program. When implemented, it will support OE by providing a viable and efficient means for managing, using, and providing access to data collected during OE program-sponsored activities. The exploitation of new data management, dissemination, analysis, and presentation techniques is a specifically stated objective within the OE program Strategic Framework[7] and supports the goal of development, integration, and application of new technologies. The strategy contained in this document supports all five of the candidate goals set forth in the Strategic Framework and will play an integral role in achieving each goal's objectives. This strategy identifies key system capabilities, alternatives architectures, and recommended approaches that will allow OE to design and implement a core capability for collecting, storing, and providing access to associated data and information products. It also involves both public and private sectors by facilitating access to exploration data. The strategy includes the following elements:

- Business drivers, goals, and objectives for managing data
- Applicable enabling technologies and standards
- Exploration data identification and processes for data exploitation
- An OE data flow model
- Metadata requirements and applicable standards
- Governing data policies and policy development needs
- Education and outreach impacts on data access processes
- High-level alternative architectures and related recommendations

This strategy has been developed through analysis of current ocean exploration activities, a review of the literature on existing and emerging ocean exploration and data management techniques, technologies, and policies, and extensive liaison with NOAA line offices, other federal agencies involved in oceanographic data collection, non-government entities within the ocean exploration stakeholder community, and experts in technologically advanced data management systems. Processes and procedures have been investigated and interpreted within the context of applicable laws, policies, and administrative rules governing the protection of intellectual property rights (IPR) and access to data collected using public resources. Architecture alternatives and a recommended high-level design for managing exploration data are also included. Lastly, a concept of operations describes some of the requirements and considerations for implementing this strategy.

## 1.3 Document Overview

Section 2 introduces governing principles for managing OE data. It lists the business drivers, goals, and objectives for managing these data. Section 3 presents the management environment within which OE data will be managed. This section discusses applicable assumptions, constraints, and enabling technologies. It outlines the OE data management process and presents a data flow model representing the path of OE data from collection to archival. It also presents applicable policy guidance, partners, and related responsibilities. Section 4 provides alternative architectures for managing OE data and includes recommendations for implementation. The appendices offer more detailed information, act as a ready reference, and support the theses presented in the main document.

# 2 DATA MANAGEMENT PRINCIPLES

NOAA and the national ocean exploration community must be prepared to embrace new principles for the dissemination of exploration data. The OE strategic goals include requirements to observe, assess, record, sample, and map the characteristics of the oceans using multiple observations over time and space. These strategic goals drive the need to adopt a supporting and robust set of goals for collecting, processing, storing, providing access to, and archiving data produced under OE sponsorship.

## 2.1 Business Driver

The OE program vision statement identifies[8] OE as the prime collection manager for data and information from ocean exploration activities and the central instrument for sharing such data and information. The goals identified in this framework that support the OE program vision are provided as follows:

- Discovery of New Ocean Resources
- Acquisition of New Knowledge about the Oceans
- Development, Integration, and Application of New Technologies
- Involvement of Stakeholders in New and Innovative Ways
- Preservation of America's Maritime Heritage

These goals place further requirements on OE to manage the collection and distribution of large volumes of ocean exploration data. The OE program vision and goals result in the identification of the following business driver, which is crucial for a successful program:

> *Collect and Distribute Ocean Exploration Information*. Collect data and information from ocean exploration activities and share this information in such a way that is available to all stakeholders, including the general public.

## 2.2 Data Management Goals

The following subsections describe goals and objectives of the data management strategy, grouped by data management topic. Each bullet presents an OE data management goal; the supporting objectives immediately follow.

### 2.2.1 Collection and Processing Goals

- Ensure that responsibilities for managing data are included in OE policy

  *Objective:* Develop a template of required data management implementation actions for cruise instructions

- Facilitate data exchange by enforcing data standards in OE policy guidance

  *Objective:* Catalog all data and events associated with OE in cruise summary reports

  *Objective:* Develop follow on action plans and "tickler" systems to retrieve data after cruises are complete

- Create and maintain metadata and perform basic quality assurance to facilitate search and retrieval of OE data either before or after delivery to storage

  *Objective:* In cooperation with the National Ocean Service (NOS) Special Projects and National Marine Sanctuaries Division, work closely with exploration data providers to ensure that the applicable standards are employed for all datasets managed by OE

  *Objective:* Provide non-proprietary metadata with associated data in a long-term, standard, stable format by employing clearly written documentation and appropriate tools

### 2.2.2 Storage Goals

- Using OE data sets, work closely with federal, state, and local agencies, academic institutions, nonprofit organizations, and the private sector to create unified, long-term databases of exploration data types to facilitate outreach and education

  *Objective:* Ensure clear roles and responsibilities for OE data management through agreements and other coordination instruments with NOAA and external organizations

  *Objective:* Organize data types into a robust catalog using stakeholder input where possible and appropriate

- Maintain OE data integrity and quality

  *Objective:* Provide quality assurance of data and information standard maintenance and security tools to promote customer knowledge of data accuracy and utility

### 2.2.3 Access Goals

- Coordinate with NESDIS to develop and maintain a process of efficient access to available OE data; ensure the quality of the data and associated metadata

  *Objective:* Allow searches of the catalog by specific mission and various thematic and temporal categories, followed by a select-and-acquire function

  *Objective:* Develop a catalog with NESDIS that allows access to data via the Internet and provides the capability for data to reside at distributed repositories to support the wide scope of OE data and information

*Objective:* Establish a robust quality assurance and metadata generation program that integrates requirements of NESDIS, principal investigators (PIs), and OE to facilitate complete and accurate access to data

- Populate and maintain databases, and provide on-line access to the public

  *Objective:* Establish a procedure for facilitating and tracking storage of and access to OE data; include a visible, tangible incentive for PIs to provide full and open access by the public to their collected data

- Facilitate timely response methodology to access data collected through OE explorations

  *Objective:* Maximize the efficiency of data access wherever possible by requiring that data management staff of OE and designated expedition data managers develop procedures with the NESDIS Data Centers—the National Oceanographic Data Center (NODC) and the National Geophysical Data Center (NGDC)—to maximize the efficiency of data access wherever possible

- Present data and data sets of different types in appropriate format for different user communities

  *Objective:* Work with NESDIS to provide data integration capabilities that combine data from different data sources; provide these data to users as new products, including new and existing techniques of visualization, geospatial information systems, analysis, and trends information

- Provide a mechanism for OE to assess, measure, and report program accomplishments

  *Objective:* Establish measures of effectiveness of data service to the public

  *Objective:* Establish visible means for monitoring PI compliance with OE data management policy guidance

### 2.2.4 Archive Goals

- Archive data for further analysis and product development by the public to help form the basis for research, outreach, and education and to facilitate public policy decisions

  *Objective:* Produce retrospective analyses and trend information

  *Objective:* Develop a state-of-the-art atlas that provides a geospatial record of OE data and identifies other supporting products for public use

  *Objective:* Prepare technical publications and papers/articles for scientific journals and meetings

- Submit copies of data and metadata to the appropriate NOAA Data Center for archival as soon as possible to ensure protection and preservation of the data

  *Objective:* Establish procedures for forecasting and tracking the transfer of data to the archive

- Archive data in a stable format on National Archives and Records Administration (NARA)-approved media and migrate to new media to meet evolving standards

*Objective:* Coordinate with NESDIS on the mechanism to establish and maintain the variety of OE data types within the archive

*Objective:* Work with NESDIS to develop an efficient video data management system

## 2.3 Distributed Data Management

The requirement for a distributed approach to data management is driven by the variety and complexity of oceanographic data types and the diversity of stakeholders. It is a primary emphasis area in managing data collected by the OE program. This approach has been incorporated throughout the data management strategy. The OE strategy is designed to take advantage of and integrate distributed object and other technologies to establish an ocean exploration data catalog. This catalog will provide electronic access to data via the Internet, allowing data to reside almost anywhere. This will expose a larger realm of data—both inside and outside of NOAA—using a modular and layered approach that best serves the customers of OE data. The catalog will provide directory-type services, like pointing to other data catalogs and repositories of exploration data. In addition to providing a data discovery service, the catalog will be designed to facilitate direct access to these other sources of data. The cataloging and access process will be compatible with existing catalogs from other NOAA data providers. The catalog will be based on standard, compliant metadata and can be linked or replicated at any number of national or international data clearinghouses.

This focus on distributed data management is consistent with recommendations made by a committee from the Computer Science and Telecommunications Board of the National Research Council (NRC) to the Library of Congress.[9] This report examined strategic directions for the application of information technology into the next decade and identified opportunities for access to and preservation of library collections using digital technologies. The report recognized that no single institution can hope to collect all or even a majority of its desired digital content and that cooperative agreements for distributed collections are essential and need to be pursued aggressively.

There are multiple technology alternatives that can be applied to the challenge of managing data within this highly distributed environment. These alternatives are discussed in Section 4.

# 3 DATA MANAGEMENT ENVIRONMENT

The environment in which OE must develop a data management capability is one of rapid change, conflicting guidance on proprietary data, and an explosion of data complexity and quantity. The growing challenges of this data environment are being addressed by an abundance of information systems solutions bounded by policies of national need, security, and proprietary rights. This section describes assumptions, constraints, enabling technologies, business processes, policies, and other factors that relate to the management of OE data.

## 3.1 Ground Rules, Assumptions, and Constraints

Ground rules, assumptions, and constraints establish the framework within which the OE data management strategy has been developed. Ground rules describe the basic choices of scope and focus made by OE. Assumptions are the underlying "truths" and predictions, taken as fact, which this data management strategy is built upon. Constraints are those legal, policy, or operational realities that further restrict scope.

### 3.1.1 Ground Rules

The following list provides the ground rules under which this data management strategy has been developed:

- In accordance with NOAA policy, OE will take advantage of the data services charter of NESDIS and partner with the national data centers, where necessary, to develop the technology and capacity to provide:
  - Long-term stewardship of data collected under federal sponsorship, including non-NOAA data collected during OE program activities
  - An environment that complies with national policy, standards, and procedures
  - Compatibility with other NOAA data management systems to ensure cost-effectiveness, efficiency, and responsiveness to the needs of the public
  - An emphasis on compatibility with non-government, public, and international partners and stakeholders to maximize the societal benefit of collected data
  - The capability to provide access to OE data sets in a timely fashion using a cost-effective process, including public on-line Internet access to these data in open standards format
  - Support for the development of a NOAA-wide video data management system that will provide OE with a coordinated, responsive method for archiving and accessing analog and digital video data
  - Cost-sharing arrangements necessary to implement the OE data management capability

  - Archival services that seek to take advantage of the NOAA emerging large-volume data archive infrastructure investment initiative—the Comprehensive Large Array-data Stewardship System (CLASS)—as part of the NOAA national infrastructure

- The OE data management strategy directly addresses digital data; physical samples, specimens, and other analog data will include digital representations in the form of metadata
- Although use of a management information system (MIS) is addressed relative to program management needs and limited access to data, it is not incorporated as a component of the architecture alternatives provided in this document
- This strategy describes a potential future data management environment and does not extend to the discovery, mining, and rescue of existing data from the vast and diverse archives that may be held by the stakeholder community
- Management of Level 4 metadata—represented by the publication of data and related results in peer-reviewed manuscripts or technical reports—is not addressed in detail within this strategy
- Due to the unique infrastructure requirements and existing or planned capabilities within NESDIS, ocean exploration data from space-based remote sensors are not included as part of the data management strategy discussed in this document

### 3.1.2 Assumptions

Throughout this document, various assumptions are made concerning the environment in which an ocean exploration data management capability would operate. The following list includes high-level assumptions that were made while identifying alternative OE data management strategies:

- Consistent with similar multidisciplinary programs, ocean exploration data will be made available in such a format that the information can be accessed by the broad user community within one fieldwork cycle—defined as one year[10]—to maintain its value to data stakeholders
- Data collected using a narrow observation strategy designed to validate research findings are considered research data, while data collected through disciplined, diverse observations with discovery as the principal focus are considered exploration data
- The process of making serendipitous discoveries during the conduct of routine, narrowly focused, and programmatically bounded oceanographic research does not constitute ocean exploration since uncovering a new discovery is not the primary intent of a researcher involved in this type of data collection activity
- In accordance with federal data management policy, OE may set user charges for data and information dissemination products at a level sufficient to recover the cost of dissemination—but no higher—and will set user charges at less than cost or eliminate them entirely where such charges would constitute a significant barrier to use by public stakeholders[11]

- In keeping with the spirit of discovery, OE data and information will be made available through diverse media—including electronic formats and the Internet—to facilitate and promote open and efficient use and exchange of these data by other government agencies and the public
- Roughly five percent of the annual OE budget will be dedicated to data management activities
- The OE budget will progressively grow over the next ten years—between five and 15 percent per year—as NOAA continues its national leadership role in exploring the oceans
- Participating PIs will continue to assume responsibilities related to the processing and quality control of raw data collected under the sponsorship of OE
- Emerging OE policy guidance to PIs concerning their responsibilities and accountability for managing collected data will be applicable to activities under both full and partial OE sponsorship
- The anticipated OE staffing plan will include manpower that can be dedicated to data management policy development, oversight, and enforcement
- With few exceptions, most ocean exploration data can be represented as a layer of information within a geographic information system (GIS) at a given point in time because of its geospatial dependence
- A wide variety of processing techniques exist that are associated with the capture, quality control, manipulation, calibration, and compression of raw data by participating PIs; while the individual techniques are beyond the scope of this strategy, their impact on the data and related format standards are addressed
- Lack of a standard, easy-to-use NOAA metadata generation tool will increase the burden on PIs to produce compliant metadata within desired turn-around times

### 3.1.3 Constraints

The data management strategy outlined in this document was developed within a set of constraints. Recognition of these constraints provided useful boundaries on strategy development and helped elucidate the potential impacts of policy, technology, and budget on eventual implementation of the strategy. Constraints with noteworthy impact on the data management strategy are included in the following list:

- OE is obligated to comply with applicable federal directives mandating the use of national standards for data and metadata format, data transfer, data ownership, and data disposition to ensure intra- and interagency leverage and recognition of public rights to these data
- While ease of public on-line access to and convergence of NESDIS data storage and data archival facilities will increase with future initiatives—such as CLASS—the combination of current national data center technologies and the unique need of the OE program to

provide a wide stakeholder community with early access to data will necessitate establishing an OE data repository

- Existing federal and NOAA policy guidance governing IPR of PIs involved in collecting OE data could result in as much as a one-year delay in availability of collected data to the public
- The implementation timeframe of an OE data management capability will be limited by the availability of applicable resources—no more than five percent of the annual OE operating budget—although cost-sharing agreements with other NOAA line offices, particularly NESDIS, will facilitate faster realization of the needed capability
- Due to the wide diversity of partnerships involved in the OE program, data management guidance to individual PIs may need to be tailored to certain partners depending on the mix of sponsors and copyright considerations
- Because of the large volume of data associated with digital video and the associated bandwidth requirements necessary to provide on-line access to these data over the Internet, the level of access will be limited to users who are collocated with the video data, with on-line users exploiting frame grabs and video clips of pertinent subject matter, as well as applicable compression techniques
- The OE objective of continually developing and integrating new technologies for exploring the oceans will result in a constantly changing subset of collected data that is not represented by the existing data model, and will require a systematic expansion and improvement of the data management capability

## 3.2 Enabling Technologies

Effective management of data collected during exploration activities will require the leveraging of new technologies under investigation by NESDIS for NOAA-wide programs, as well as the incorporation of new technologies specific to the needs of these unique data.

### 3.2.1 Applicable Technologies and Standards

Database management system (DBMS) technology is applicable throughout the data management process, from collection on site through the storage, access, and archiving of exploration data. In particular, the central repository for data available to users through an on-line access capability is expected to require at least one mature, fully featured DBMS. While this repository could also act as the archive for these data, it is likely that the technology applied to data archiving will emphasize the long-term safety and stewardship of these data and the capacity for accommodating a large data volume at the expense of on-line access capabilities. A relational or object-oriented DBMS would satisfy the basic data management requirements of persistence, secondary storage management, concurrency,

recovery, and *ad hoc* query. The specific mix of database technologies most appropriate for OE data will result from the systems engineering and design process. The following attributes allow a comparison between these types of data repositories:

- Relational DBMS technologies are more mature than those associated with object-oriented DBMS
- An object-oriented DBMS offers the developer advanced features that support modeling complex objects and relationships, including the ability to give users access to data in an open standard format even though the data are stored in a variety of native formats. The data management process being incorporated at the National Coastal Data Development Center (NCDDC)[12] is an example of the use of object-oriented technologies to provide a broad spectrum of on-line users access to oceanographic data in open standard formats
- A relational DBMS conforms closely to a common, open model; object-oriented DBMS models are less consistent, resulting in applications that have the potential to be more closely tied to proprietary solutions
- A relational DBMS is well suited to the requirements of handling metadata

Standards will be required to coordinate the data across multiple databases. The American National Standards Institute (ANSI) adopted Structured Query Language (SQL), a common database query language, as an industry standard in 1986. The International Organization for Standardization (ISO) has since formally recognized SQL as a standard,[13] as has the International Electrotechnical Commission (IEC) [ISO-9075] and the federal government in its Federal Information Processing Standards (FIPS). SQL provides portability of database definitions and database application programs among conforming implementations. It is appropriate to use the SQL standard in all cases where there is an interchange of database information between systems. The SQL definition language may be used to interchange database definitions and application-specific views. The SQL data manipulation language also provides data operations that make it possible to interchange complete application programs.

Remote Database Access (RDA) is a communications protocol for remote database access that has been adopted as an ISO/IEC standard. RDA provides standard protocols for establishing remote connections between a database client and a database server. The client acts on behalf of an application program, while the server interfaces to a process controlling

data transfers. RDA promotes the interconnection of database applications in a multi-vendor environment.[14]

The Object Data Management Group (ODMG) is a consortium of vendors and interested parties who collaborate to develop and promote standards for object storage. The current standard, ODMG Release 2.0, is intended to ensure the portability of applications across different database management systems. The ODMG specification is a set of components that include an object model, an object definition language, an object query language, and bindings to languages such as Java, C++, and Smalltalk.

The application of GIS technology is relevant to data that can be referenced using geospatial coordinates, making it particularly applicable to ocean exploration data. The strategy for managing these data includes the use of GIS to provide sophisticated and enlightening access services to a wide range of data users. A GIS is both a database system with specific capabilities for spatially referenced data and a set of operations for working with these data. A GIS consists of a system of hardware and software that combines graphics and databases to generate layered maps and reports, enabling users to collect, manage, and interpret information in a planned and systematic manner. These users may need information from many different sources in many different forms to perform scientific analyses. Such a user may have a GIS or may use GIS services provided by NOAA. At a minimum, the data must be available in one or more of the standard formats compatible with the user's software applications (e.g., GIS or Web browser). The standards published by the OpenGIS™ Consortium[15] allow data users to employ several different and popular commercial GIS tools. Map standards within any GIS must meet National Map Accuracy Standards (NMAS) established by the U.S. Bureau of the Budget in 1941. These map standards have been revised many times, through the current version employed by the U.S. Geological Survey (USGS).

The Spatial Data Transfer Standard (SDTS) is a mechanism for archiving and transferring spatial data (including metadata) between dissimilar computer systems. SDTS specifies exchange constructs, such as format, structure and content for spatially referenced vector and

raster (including gridded) data. SDTS was approved initially as FIPS Publication 173 (also by the Federal Geographic Data Committee (FGDC) as STD-002), but has been encompassed as a federal data transfer standard within the ANSI International Committee for Information Technology Standards (NCITS) 320-1998. Other spatial data standards that have gained acceptance in the international community include the Digital Geographic Information Exchange Standard (DIGEST) and the Spatial Archive Interchange Format (SAIF).

The Intergovernmental Oceanographic Commission (IOC) has recognized the Marine Geophysical Data Exchange Format (MGD77) as an accepted standard for international data exchange; it has been translated into French, Japanese, and Russian. The digital format for MGD77 is an exchange format originally developed in 1977 that is suitable for marine geophysical data (bathymetry, magnetics, and gravity). It was intended to be used for the transmission of data to and from a data center and may be useful for the interchange of data between marine institutions. NGDC has distributed MGD77 as its standard exchange format since late 1977.

Table 3-1 provides a summary of the high-level technical problems associated with managing ocean exploration data, desired capabilities, and related candidate technologies for addressing these needs. Challenges specifically associated with managing digital imagery and video are addressed in the following section. Metadata technologies are discussed in Section 3.2.3. Innovative ship-to-shore radio frequency (RF) communications links such as the experimental high-speed wireless link being installed on the Research Vessel (RV) Roger Revelle at the Scripps Institution of Oceanography[16] have the potential to impact exploration data management, particularly in support of real-time participation of PIs and data users from remote locations during data collection. These communications technologies are briefly addressed in this strategy. From a data-flow perspective, the impact of these communications simply changes the physical location where data collection and processing takes place. The functional flow of these exploration data are essentially unchanged.

**Table 3-1. Technical Problems Associated with Data Management**

| Problem Area | Capability | Candidate Technologies |
|---|---|---|
| On-ship data management | • Metadata collection | • Metadata tools<br>• Barcode generators, scanners for physical media |
| Ship-to-shore data transfer | • Cross-platform data transfer | • High-speed RF communications |
| Central repository storage and retrieval of data and metadata | • Data storage and on-line access in open standard formats | • DBMS<br>• Object-Oriented DBMS<br>• Open GIS frameworks |
| Archive and retrieval from archive | • Large volume archive | • DBMS<br>• Massive storage devices |
| Metadata standards for central repository and archive | • Universal access to metadata and data | • CSDGM (FGDC-STD-001-1998)<br>• NBII<br>• XML |
| Metadata standards for externally managed data | • Universal access to metadata and data | • XML |
| Central repository for video data archival and access | • Coordinated process for handling and providing access to video data | • Video archive facility<br>• Logging, indexing, and annotation software<br>• Data retrieval system |

### 3.2.2 Video Data Management

The use of digital imagery by explorers and researchers as a method of capturing, processing, and storing unique information about the ocean environment is growing as supporting technology has increased capabilities and reduced costs. Still digital imagery and digital video are quickly surpassing traditional cameras that use film to record and produce images. Advantages include scalability, adaptability to varying lighting conditions and wavelengths of interest, ability to reuse memory, and capability to quickly distribute results electronically. Use of this technology has resulted in new challenges related to the process of storing, editing, annotating, archiving, and providing access to large and increasing volumes of data. Digital cameras are able to capture images at a resolution of three or more megapixels. For moving images, a high-speed external bus is usually required to achieve desired transfer speeds and to effectively use imagery applications. The processing of digital video imagery includes creating time-referenced content annotations—metadata—either at regular time intervals, such as a frame-by-frame basis, or coincident with subject matter of interest as determined by an observer or editor. Software must be employed with advanced search-and-retrieval technology

to facilitate fast and accurate browse, search, and preview of video source material. Similar technology must be employed by production personnel in order to search vast archives for relevant footage, automatically capture video clips, browse storyboards, catalog content, add text and voice, create rough cuts, create edit decision lists (EDLs) for further production, and post the resulting products to other media, such as an Internet web server, for access by the user community. Emerging operational considerations related to the collection of oceanographic data include the optimal location and timing for image or video annotation—at the scene or during post-analysis—and the exploitation of sophisticated digital video editing software packages for annotating data with metadata that are more appropriate for use by experienced editors at a centralized location. Examples of video data management technologies and processes that could be applied to OE data are in Appendix A.

### 3.2.3 Metadata

Data that are used to describe the content, representation, structure, and context of some well-defined sets of observational data are called metadata. Metadata are required to facilitate the identification and acquisition of companion data, determine the data's suitability for meeting a specific objective, and support additional processing, analysis, and use in numerical models. Documentation of metadata is vital to a dataset's accessibility and longevity for reuse. Without this documentation, other scientists cannot know what suitable datasets already exist to answer a particular research question. Additionally, as time passes, the data become unusable as information about the data is lost.[17] Information loss hampers data sharing and is ultimately detrimental to science as the data themselves become unusable.

Metadata are generally considered to be the information necessary for someone who is not previously acquainted with the data set to make full and accurate use of that data set. At a minimum, the metadata associated with a data set must provide a consistent framework that accomplishes the following objectives:

- Permits assessment of the applicability of the data set to the question or problem at hand
- Supports assessment of the quality and accuracy of the data set
- Provides all necessary information to permit a user to access and understand the values in a data set

- Permits the assignment of correct physical units to the values
- Supports the translation of logical concepts and terminology among communities
- Supports the exchange of data stored in differing physical formats

Four levels of metadata have been defined[18] that have applications at various stages of the data management process, as listed in Table 3-2. Level 1 metadata consists of the basic description of the field program collecting the data, including location, program dates, data types, collecting institutions, collecting vessel, and participating investigators. Level 2 metadata consists of a digital summary report and data inventory that is created shortly after completion of the data collection process. For oceanographic applications, this level of metadata typically takes the form of a cruise report. An international standard known as a Cruise Summary Report (CSR)—formerly known as the Report of Observations/Samples collected by Oceanographic Programmes (ROSCOP)—has been in existence for more than 20 years and has been used with some success by a number of data centers, notably the International Council for the Exploration of the Sea (ICES). The CSR provides a brief and informative summary of data collected on a cruise, including the types of data that were collected, in what amount, by whom, in what area, and when they were collected. The resulting information is available to scientists and planners through world and national data centers. The ICES CSR database is supported by both retrieval and form entry software, which are suitable for use on most personal computers; in addition, the CSR data entry software is available for download from ICES.[19] Level 3 metadata consists of data object and access information, including data formats, quality control, processing, and any elements necessary to describe subsequent changes in the content, format, and accessibility of the companion data. During the lifetime of a data set, many modifications to its Level 3 metadata can be expected as it is manipulated and combined with other data to derive additional information in the form of companion data sets. Finally, Level 4 metadata describes the process of formal publication of results derived from the data. Further discussion of Level 4 metadata outside the scope of this data management strategy.

There are several basic classes of information that should be provided as metadata components of ocean exploration data regardless of the source:[20]

**Table 3-2. Metadata Classification Levels**

| Classification Level | Description |
|---|---|
| Level 1 | Basic description of the field program |
| Level 2 | Digital summary report and data inventory (e.g., CSR) |
| Level 3 | Data object and access information |
| Level 4 | Formal publication of results derived from data |

- Type (physical, biological, chemical, geological, or archeological)
- Volume
- Discipline (tidal, waves, currents, ocean color, plankton, cetaceans, cephalopods, sediments, mineral deposits, coral reefs, wrecks, etc.)
- Data class (text, numeric, acoustic, image, video, etc.)
- Observation regime (deep ocean, coastal, surface, etc.)

These classes of information must be defined through established policy prior to data collection activities so that PIs have a consistent set of guidelines and minimum standards for metadata content and quality.

For each data set collected, NOAA-funded investigators are obliged to submit metadata as soon as possible after data collection.[21] NOAA line offices typically require these metadata in a predetermined, standardized format within 60 to 90 days of data collection. Under Executive Order 12906, all federal agencies and organizations receiving federal funds must document geospatial data using the FGDC Content Standard for Digital Geospatial Metadata (CSDGM).[22] The National Biological Information Infrastructure (NBII) Content Standard for Biological Information, which is compatible with the CSDGM, has become accepted as a mature, universal, and complementary metadata standard for biological data. These metadata standards have been adopted by NOAA and continue to be emphasized through active participation with FGDC in the National Spatial Data Infrastructure (NSDI). With regard to NOAA participation, it is important to note that formal NOAA policy guidance regarding metadata standards and compliance with these standards does not exist.[23] Metadata policies are discussed in Section 3.5.

Metadata can be created using tailored software or by using a text editor. A wide variety of tools—with varying degrees of functionality—are available to those charged with metadata

development.[24] Regardless of the method used to create the metadata, the current infrastructure within NOAA requires the American Standard Code for Information Interchange (ASCII) format for incorporating metadata. There has been considerable interest within the Internet community on the use of the Extensible Markup Language (XML), a Worldwide Web Consortium (W3C) recommendation for the packaging of structured information. XML provides a reference framework for encoding the nested data structures that occur in metadata and a means (validating parsers) to test them. A draft-encoding standard for digital geospatial metadata—in effect for several years—defines a formal XML encoding of FGDC metadata. The encoding is structurally enforced using a reference file known as a Document Type Declaration (DTD) that is hosted on the FGDC website.[25]

Two issues concerning metadata are of the most critical importance for developing a strategy to manage ocean exploration: (1) establishment of policy guidance and (2) provision of guidelines for participating investigators to ensure that responsibilities and accountability for creating, delivering, and updating metadata are clear and consistent. Recommendations for managing metadata are included in Section 4.

### 3.2.4 Management Information Systems

A MIS is generally defined as the complement of people, machines, and procedures that develop the right information and communicate it to the right managers at the right time.[26] For the purposes of this strategy, a MIS for OE would provide access to programmatic, historical, and budgetary information necessary for OE to effectively manage the OE program. An OE MIS would include but not be limited to the following: contact and background information on former and current partners and collaborators; the capability to receive and manage proposals from prospective PIs using established criteria; access to Level 2 metadata from completed exploration activities and a select subset of data—such as descriptive imagery or video clips—associated with those activities; and a means for tracking and reporting operations and costs that support the assessment of completed expeditions, the enforcement of commitments in investigator awards, and the measurement of the nature, volume, and success of exploration achievements. By necessity, an OE MIS would not be designed to provide capabilities for storage and external access to the large volumes of

fundamental data collected during an exploration campaign, and as such, the role of the OE MIS is only briefly addressed by this data management strategy. Two candidate MIS solutions under development within NOAA are discussed in Appendix B.

## 3.3 Ocean Exploration Data Management Process

Managing enormous amounts of oceanographic data gathered from different instruments and sensors via a wide variety of exploration activities is a daunting task. The ocean is a turbulent fluid that is constantly changing over many spatial and temporal scales. Each data item gathered possesses unique information as long as it is accurate, corresponds to a different quantity, is obtained from a different time and place, and cannot be accurately computed from the existing data. Observed oceanographic data are largely non-redundant in nature. Each observation is a unique datum that—due to the passage of time alone—cannot be replicated. As such, all non-redundant data will be needed by future generations. Non-redundant data that are destroyed cannot be recovered since the oceans are dynamic and past observations are nearly impossible to reconstruct from other data.[27]

Exploration data are typically generated by a diverse group of investigators who use a wide range of techniques to gather, archive, quality control, and distribute this data. There are also times when investigators, for various reasons, cannot or choose not to distribute the data at all. There are times when the same physical attributes, such as temperature or salinity, might be gathered by different instruments using different data formats and data quality standards before it is stored for later use. Some of these collection methods are highly sophisticated, whereas others may be too crude to merit replication by other researchers at a later time. Thus, some form of data management process needs to be established that will enable a variety of users to access the vast majority of data, with a measure of the fidelity of the data, in a standardized format for further analysis.

The planning, implementation, and maintenance of an effective mechanism for long-term archiving of observational data sets must address three critical issues: storage management, accessibility, and "assessability." Storage management focuses on various aspects of archiving, including the reliable storage of data for long periods of time, transfer of data from

old to new storage technology, physical data distribution to accommodate institutional policies regarding custodianship or the physical limitations of an institution, and retrieval performance requirements. Accessibility concerns include the provision of capabilities that provide a model of interaction and a mechanism for accepting input from a user on information needs, locate all data relevant to those needs, and retrieve, package, and deliver the needed data to the user. Assessability permits the user to clearly determine the significance, relevance, fidelity, and quality of the data.

A 1995 report of the NRC addressed issues concerning the preservation of scientific data on the physical universe. Six of the conclusions reported relate to the processes of managing data and metadata.[28]

- Effective archiving needs to be begin whenever a decision to collect data is made
- Originators of the data should prepare them initially so that they can be archived or passed on without significant additional processing
- The greatest barrier to contemporary and future use of scientific data by other researchers, policymakers, educators, and the general public is lack of adequate documentation
- A data set without metadata or with metadata that do not support effective access and assessment of data lineage and quality has little long-term use
- For data sets of modest volume, the major problem is completeness of the metadata, rather than archiving cost, longevity of media, or maintenance of data holdings
- Lack of effective policies, procedures, and technical infrastructure—rather than technology—is the primary constraint in establishing an effective metadata mechanism

This suite of conclusions led the committee to recommend that "adequacy of documentation" should be a critical evaluation criterion for data set retention.

A data management workshop for Marine Geology and Geophysics (MG&G) was held in 2001 through the National Science Foundation (NSF) and the Office of Naval Research (ONR).[29] The recommendations resulting from this workshop represent a unique and applicable set of implementation recommendations for managing a key subset of oceanographic data, many of which have been applied to this strategy. The list of the recommendations resulting from this workshop is in Appendix C.

### 3.3.1 Data Types

There are a variety of data types that will be produced by the OE program. These data types are described by multiple data formats and may include an accompanying taxonomic scheme to provide insight into the data's attributes. The expected volume of data has an impact on the strategy for managing them and will drive the design of the supporting system. A data flow model provides a foundation for follow-on data management systems engineering efforts and describes the lifecycle of ocean exploration data.

#### 3.3.1.1 Data Formats

The OE program will collect a variety of data in digital and non-digital formats. These data sets include point, line, grid, and other spatial observations and measurements, acoustic observations and time series, geological observations, biological and geological samples, chemical analyses, archeological artifacts, and considerable volumes of imagery and video. Data formats will be driven by the requirements of participation PIs during collection and guided by OE policy during storage and archival to meet the data access needs of ocean exploration stakeholders. A discussion of the common data formats and associated media used to collect these data during ocean exploration activities is contained in Appendix D.

#### 3.3.1.2 Taxonomy

Implementing a data management capability to support the OE program can be enhanced by developing a taxonomic scheme for exploration data. This taxonomy would provide systems engineers with insight into the scope and attributes of data that will be stored, manipulated, and accessed by the system. It can also provide the foundation for the development of a data model that aids the development of data and information management processes and helps the supporting system maximize its potential. A developmental scientific information model associated with the NOAA VENTS program at the Pacific Marine Environmental Laboratory (PMEL) is one example of a data modeling process with an emphasis on use by a GIS that could be applied to the OE program.[30] The program that is the subject of this approach has several similarities with the OE program:

- They are both interdisciplinary initiatives that encompass a wide range of disciplines, including geophysics, geology, physical oceanography, chemistry, and biology

- Each program has an objective of implementing a data management system that integrates storage, access, and archival functions
- Both programs recognize the geospatial attributes of the vast majority of cognizant data and the potential offered by a GIS to provide unique interpretive and product generation capabilities

Appendix E provides a GIS-based taxonomy template for exploration data that could form a basis for an eventual data model. It includes GIS topology and sample, high-level attributes for each data type. The topology represents the spatial relationship between connecting or adjacent features in a GIS coverage. This taxonomy could be augmented to meet implementation needs for a data management system through the creation of an ocean exploration data model. By necessity, such a model would be a living document, continually modified and expanded as new data types and definitions are encompassed by the OE program and new technologies are integrated that produce fundamentally new data and information.

Due to the anticipated dependence of ocean exploration on imagery and video, embarking on a coordinated data modeling effort would also help users interpret video data. The model could be used to define language syntax for use by a subject matter expert involved in the annotation of video data, similar to a process that has been used at the Monterey Bay Aquarium Research Institute (MBARI).[31] A data model could also provide insight into additional metadata and lineage requirements of selected data. Camera tow and seafloor mooring GIS coverages, carrying an extensive number of attributes, require more detail than coverages such as bathymetry that have a more easily represented topology in the GIS.[32]

#### 3.3.1.3 Storage Volume

The OE program will produce a large volume of digital data. Of these data, the requirements related to the management of video data can be expected to drive video data volumes that reach 90 percent of the total storage volume demanded by the OE program. Table 3-3 provides approximate volumes of video and non-video data collected during the inaugural 2001 OE campaign. These values were obtained by reviewing cruise summary information for the 2001 OE campaign and direct contact with chief scientists, PIs, and data managers involved in the expeditions. Given the increase in the OE budget in 2002, a concurrent increase in the volume of collected data is suggested. New developments in sonar, remotely operated vehicles

(ROVs), autonomous underwater vehicles (AUVs), and submersibles also promise to accelerate the rate at which data are collected. Resulting increases in the rate of new discoveries may be expected to drive demand for access to this data.

**Table 3-3. Volumes of Data Collected during Inaugural 2001 OE Campaign**

| Mission | Data Collected (Gigabytes) | |
|---|---|---|
| | Non-Video | Video |
| Lewis & Clark Legacy | 2 | 850 |
| Islands in the Stream | 1 | 50 |
| Deep East Expedition | 3 | 400 |
| Sound in the Sea | 50 | — |
| Davidson Seamount | <1 | 1,030 |
| Thunder Bay | 50 | — |
| Preserving the USS Monitor | 10 | 550 |
| Total | 116 | 2,880 |
| **Total Non-Video and Video** | **2,996** | |

Table 3-4 summarizes the projected data storage and archival requirements for an ocean exploration data management system. It recognizes the variability among individual expeditions resulting from different suites of sensors and instruments and factors in the steady OE program growth discussed in Section 3.1.2.

**Table 3-4. 2002 Projected Data Storage and Archival Requirements**

| Timeframe | Non-Video Digital Data | Video Digital Data | Archive |
|---|---|---|---|
| Single Expedition | 1-100 GB | 50-1,000 GB | 1 TB |
| 1 year | 500 GB | 5.5 TB | 6 TB |
| 10 year projection | 8 TB | 80 TB | 88 TB |

The estimates for digital video in Table 3-4 do not take into account significant changes in video technologies that may be incorporated into exploration activities. If OE program participants transition to increased use of higher-resolution media and techniques for video data collection—such as high definition television (HDTV) format—for a significant subset of collected video data, storage and archival requirements will expand accordingly.

### 3.3.2 Data Flow Model

The collection and distribution of ocean exploration information is a challenging task for OE to undertake. It encompasses many legal, cultural, operational, and technical issues. OE will

pursue a multifaceted approach to successfully meet these challenges. The OE approach will be applied to two important data management concepts: central-versus-distributed data storage and operational-versus-archival data storage. Centrally storing all ocean exploration data is not pragmatic given the variety of challenges. Rather, OE data management strategy will make use of both approaches. That is, OE will centrally store all collected data for which the PI is unwilling or unable to host the data. At the same time, OE will assure access to data distributed among various investigators. Likewise, operationally storing all data—defined as storing data in a manner that allows it to be quickly accessed—is not practical given the volume of data that will be collected. Therefore, the OE data management strategy will include both operational data and archived data—data that does not need be immediately accessible but must be accessible within specified time limits.

For the purpose of modeling the lifecycle flow of OE data, the management process can be decomposed into five functional areas or phases, as illustrated in Figure 3-1.

- *Data Collection*. Raw exploration data and associated metadata are collected during ocean exploration activities "on location" and stored as raw mission data and associated raw mission metadata
- *Data Processing*. On the exploration platform, raw mission data are augmented through the examination and quantification of collected physical specimens, recording of metadata, and other preliminary processing steps required to produce quality controlled and complete datasets
- *Data Storage*. At the end of each expedition, all collected data are transferred from the data collection platform to operational storage facilities—either investigator host site databases or the OE central repository—and identified by metadata entered into the OE central catalog
- *Data Access*. Users access OE data and associated products through the OE central catalog by searching metadata that describes OE data location and characteristics
- *Data Archiving*. All OE data and associated metadata are permanently archived to support perpetual maintenance of the data and to guarantee access to historical data
- Figure 3-2 (foldout inside back cover) provides a detailed diagram of the data flow model. It reflects the five functional phases and includes processes, decision points, entities, connections, and the flow of data and metadata between the functional areas. The collection, processing, and storage phases are decomposed into three levels of detail that are reflected in the numbering scheme on symbol labels. The access and archive phases are presented at two

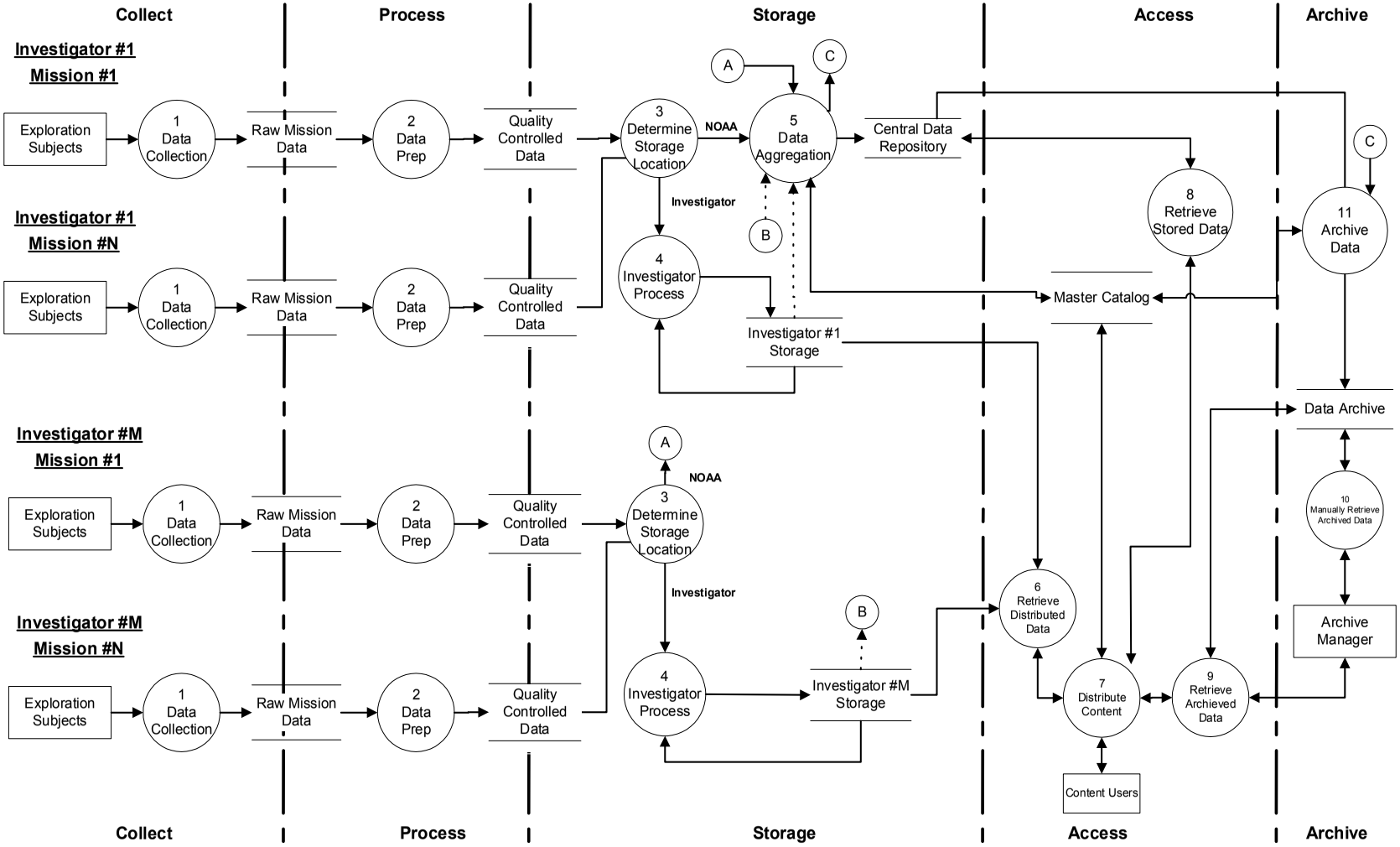



**Figure 3-1. Functional Areas of Data Management Process**

levels of detail. Figure 3-2 provides a graphical depiction of the numerous input and output functions that occur both simultaneously and sequentially during the lifecycle of OE data. It is important to recognize that the OE data management process embodies many sources of data and many different data users. Consequently, the OE data management process represented by this model has been designed to collect and provide access to data by a variety of stakeholders. The following subsections provide additional details on the processes and key decisions within each of the five phases.

#### 3.3.2.1 Collection Phase

Four entities, four processes, and three resulting data states are represented in this phase. The entities include physical samples, sources of analog and digital exploration data, and PIs and instrumentation that act as sources of Level 1 data and metadata. The collection functionality pertains to the act of collecting exploration information—data and metadata—during an expedition. Digital data are data recorded in a digital medium, while analog data are stored based on the needs of the data type. The path describing the flow of data for physical samples leads to digital data through laboratory analysis, but has been identified separately in the model to point out the different collection process for this type of data. The collection phase ends with exploration data residing in one of three data states:

- *Raw Mission Data: Digital Exploration Data.* Digital data that has been collected through the exploration process but has not yet been through a conversion, calibration, or quality assurance process
- *Raw Mission Data: Analog Exploration Data.* Analog data that has been collected through the exploration process but has not yet been through a conversion, calibration, or quality assurance process
- *Raw Mission Data: Metadata*. Level 1 metadata that are recorded during the data collection process; may include descriptions of the method(s) used and information about the environment and location where the data were collected

#### 3.3.2.2 Processing Phase

This phase of the data flow model includes one decision point, eight processes, three resulting data states, and a connection to the OE Master Catalog. There are two data flow model entities in the process phase. The processing functionality describes the application of quality control, calibration, and conversion procedures applied to raw mission data and metadata. PIs apply a variety of quality control mechanisms and calibration data to raw analog and digital

exploration data. Data may be converted from proprietary formats provided by collection instruments into formats that meet the needs of the anticipated users. During this process, a PI may convert some of the analog data into a digital medium. Metadata undergo a similar quality control process in which bad metadata elements are removed. Added details—both process information and environmental information—are added to the metadata during this phase in addition to the error correction. Once the metadata have been through the quality control process, they exist in a state that satisfies initial OE requirements for Level 3 metadata. At the end of the process phase, the analog and digital exploration data and metadata exist in three data states:

- *Quality Controlled Mission Data: Digital Exploration Data.* These are raw digital mission data that have been through a quality control, calibration, or conversion process
- *Quality Controlled Mission Data: Analog Exploration Data.* These are raw analog mission data that have been through a quality control, calibration, or conversion process
- *Quality Controlled Mission Data: Metadata*. These are metadata that have been through a quality control process and have additional detailed information (process and environment) satisfying initial Level 3 metadata requirements

#### 3.3.2.3 Storage Phase

The storage phase is accompanied by the most detailed representation among the five functional phases. This detail is a result of the data flow model accommodating both centralized and distributed storage of exploration data. This phase includes 14 decision points, 19 processes, five resulting data states, and multiple connections within the storage phase and to the access and archive phases.

Once data and metadata have completed the processing phase, they may be stored using a centralized strategy in a central repository, in investigator storage using a distributed strategy, or in the data archive. These storage locations may hold data that are accessible by the public and data that are non-accessible due to IPR or other considerations. All metadata are accessible via the OE Master Catalog. Functionally, this catalog describes the data that exists and where they are located. When new data are added or existing data are updated, the catalog is also updated. Also, the OE Master Catalog is updated as metadata in distributed locations are modified. If the data will be stored in a central repository, the data path depends on the data type. In the case of analog data, digitized samples of these data join other digital data in this

phase, while archival of analog data occurs as a separate process and connects either to the archive phase or to a process outside of the model. Digital data are sent to the archive phase as they become available. For digital data, a decision point on format conversion—that is, any additional application of quality assurance techniques, compression, information extraction, additional application of calibration data, or file conversion to comply with OE data standards—is reached, and once any format conversion has been completed, the digital data are examined to see if additional metadata are needed. Metadata are updated as necessary, and the OE Master Catalog is updated to reflect the changes. Accompanying Level 3 metadata are always provided to the OE Master Catalog, regardless of the physical location of the associated data. The associated data are forwarded to the central repository for operational use. Periodically, the operational data in the central repository are reviewed as directed by OE to determine whether continued operational access by data users is warranted. If a decision is made to remove these data from the operational repository, the data in question will be removed and the associated metadata in the central catalog will be updated to reflect this change.

If the data are stored in investigator storage using a distributed storage strategy, the paths for digital data and analog data are similar. OE policy guidance based on IPR and sponsorship will determine whether the government has rights to release data for public access. An additional decision point provides for the application of this policy to additional data sets derived from the original data by the PI. Digital data are formatted by the PI to comply with standards set by the PI's organization. The data are examined to see if additional metadata are required. If required, updated metadata are reflected in the local catalog and are also provided to the OE Master Catalog. Once this step is completed, the data are stored in investigator storage. Periodically, the PI will review these data in accordance with established OE policy to determine whether they should remain in the local repository. If removal is warranted, the data are removed and processed in the same manner as data stored in the central repository. Due to IPR issues, some of the distributed digital data may not be immediately accessible to public users. Technical issues may prevent analog data from being accessible except via physical transfer. The PI will provide a process to manually retrieve these data for a content user when appropriate.

Accurate metadata are fundamental to the successful management of ocean exploration data. OE policy guidance must include direction concerning adherence to metadata standards, frequency and content of updates, timeliness of submission, and enforcement procedures. In general, any data collected under federal sponsorship must be made available to the public. Collaboration with other organizations and industry and the incorporation of proprietary tools and methods will add complexity to this requirement for public access. No single policy guideline will satisfy every OE–investigator relationship. In addition, rigid guidance on public release may discourage the formation of valuable partnerships and alliances. Before a grant is awarded, OE must ensure that expectations regarding IPR are understood. Contract arrangements with PIs must clearly delineate this guidance and identify the source and level of resources required by the PI to comply with OE policy.

The conclusion of the storage phase is represented by data and metadata residing in the following data states:

- *Central Data Repository: Exploration Data*. Digital data that are stored in a central repository sponsored by OE
- *OE Master Catalog*. Also called the OE central catalog within this strategy, it is the directory for all ocean exploration data and is populated with associated metadata
- *Investigator Storage: Analog Exploration Data*. Analog data that are stored within one of the distributed investigator storage facilities
- *Investigator Storage: Digital Exploration Data*. Digital data that are stored within one of the distributed investigator storage facilities
- *Investigator Storage: Metadata.* Metadata that are stored within one of the distributed investigator-maintained catalogs

#### 3.3.2.4 Access Phase

The access phase describes the flow of ocean exploration data to the variety of users. It includes two decision points, 26 processes, one entity representing the data content user, and multiple connections to points within the process and to the storage and archive phases. The access functionalities of *search*, *browse*, *create new products,* and *request content from PI* are available to the content user. The browse function, similar to browsing a table of contents or index, allows the user to view contents of metadata. The search function allows the content user to search metadata for specific attributes using search capabilities provided with the OE

Master Catalog. Once the user has located desired data using search or browse, the data may be viewed or accessed on-line if the data are accessible. The acquisition of non-accessible data is similar for investigator storage and archival. In the case of non-accessible data in investigator storage, investigator contact information will be displayed to the content user. The content user would then need to contact the PI directly to make arrangements for acquiring the desired data. In the case of non-accessible data stored in the archive, the content user will exploit the access process offered by the archive facility. In cases where data are accessible, the content user may create new products from existing datasets. For both data and products in distributed investigator storage, a function is included within the data flow model that provides for on-line conversion to a standardized content format. This function accommodates the spatial data translation capability described in Section 4.1.1.

#### 3.3.2.5 Archive Phase

The archive phase includes six processes, two final data states, and one entity representing an archive manager. The exploration data are examined and the accompanying metadata are updated as necessary. These updated metadata are also reflected in the OE Master Catalog. When a user submits a request for archived data, the archive manager processes the request and distributes the data to the content user. If the archive manager offers an automated process that makes the data accessible, the content user may retrieve the data from the archive facility using the on-line process. Archive policies will be established by the NOAA Data Centers with archiving responsibility. There are two final data states in the data flow model:

- *Data Archive: Data.* Data that are resident in the NOAA archive
- *Data Archive: Metadata.* Metadata that describe the data resident in the archive

## 3.4 Data Policies

New technologies such as the Internet have created a growing public constituency and facilitate an increasing demand for data and information. This constituency extends beyond the research community to include commercial entities, policymakers, educators, nonprofit organizations, the general public, and the international community. The expected demand for ocean exploration data and information from education and outreach stakeholders represents a requirement for access to OE data that is unique within NOAA and approximated by exploration initiatives within NASA. As a result, appropriate federal regulations and policies are critical to ensuring that data are collected efficiently, preserved for posterity, and made

available for the benefit of the widest possible user community within the boundaries of IPR and commercial interests.

Advocacy groups with an interest in balancing the benefits of full disclosure of scientific data with protection of IPR to these data have sought guidance for organizations and individuals to help them evaluate legislative proposals that affect the use of scientific databases. Examples include the set of principles proposed in 1997 to the World Intellectual Property Organization (WIPO) by the ad-hoc Group on Data and Information of the International Council for Science (ICSU) and the Committee on Data for Science and Technology (CODATA).[33] These principles have been interpreted and restated here in the context of ocean exploration. They represent a foundation upon which regulatory bodies should base policy decisions that govern access to ocean exploration data.

- *Ocean exploration is an investment in the public interest*. Through discovery, education, and outreach, explorers foster the creation and dissemination of knowledge. This can have profound effects on the well being of people and the economies of the world. Exploration of the oceans is a critical public investment in our future, a resource with extraordinary dividends.
- *Ocean exploration relies on full and open access to data*. Both science and the public are well served by a system of scholarly research and communication with minimal constraints on the availability of data for further analysis. The practice of full and open access to data has led to profound discoveries, breakthroughs in scientific understanding, and benefits to economic and public policy interests.
- *A market model for access to ocean exploration data is unsuitable for research, education, and outreach*. Science is a cooperative, rather than a competitive, enterprise. No individual, institution, or country can collect all the ocean exploration data it needs to address important issues. Thus, practices that encourage data sharing are necessary to advance science and to achieve resulting societal benefits. If costs for access to these data are prohibitively high, the negative impact on the public is the same as if access were legally denied.
- *The interests of ocean exploration data owners must be balanced with society's need for open exchange of information*. Given the substantial investment in data collection and its importance to society, it is equally important that data are exploited to the maximum extent possible. Policy guidance and attitudes among participants within the oceanographic community should foster a balance between individual rights to data and the public good of shared data.

U.S. federal policy guidance and protocols for data ownership, access, storage, and liability are set forth in public law and executive order, transmitted in Office of Management and

Budget (OMB) directives concerning the management of federal information resources (OMB Circular A-130) and standards for the administration of grants to and agreements with institutions of higher education and other non-profit organizations (OMB Circular A-110). This latter directive, having recently undergone revisions with strong comment from the scientific community[34], defines the phrase "research data" as recorded factual material commonly accepted in the scientific community as necessary to validate research findings. The definition allows the identification of data that are subject to Freedom of Information Act (FOIA) requests. This directive also lists categories of data that are exempt from this definition and associated FOIA requests, including preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, physical objects such as laboratory samples, trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, similar information protected under law, and data that require protection under privacy act regulations.

Consistent with the definition of exploration in the *Frontier Report* and the assumptions used in developing this strategy, an explorer is distinguished from a researcher by virtue of the fact that the explorer, while dedicated to the conduct of exploration, is not confined to a narrow observing strategy. The boundary between these functional labels is important but frequently indistinct, recognizing the fact that a vast number of serendipitous discoveries occur during the course of scientific research. A researcher conducting research (i.e., not focused on exploration) is likely to make discoveries whether or not the researcher is using data collected during ongoing or prior exploration activities. On the other hand, a researcher might plan for and dedicate a percentage of time and resources to exploration activities during a research mission if approved by the applicable sponsor. Since, by convention, exploration data are not collected as a means to validate research findings, they do not fall within the definition of research data in Circular A-110 and thus the associated caveats related to research data (e.g., response to FOIA requests) do not appear to apply. If discoveries are made during the course of dedicated exploration activities (e.g., those likely to be sponsored by OE or other exploration-focused organizations), Circular A-110 provides the federal government with the right to obtain, reproduce, publish, or otherwise use these data, as well as authorize others to receive and use these data for federal purposes. In

exercising this right, NOAA and OE have an obligation to avoid creating situations that could lead to a misinterpretation of exploration data by the public, such as a premature release of data lacking appropriate application of quality control techniques. Data associated with serendipitous discoveries made during research activities are not considered exploration data, meet the Circular A-110 definition of research data, and the associated caveats apply.

Application of federal guidelines to specific OE activities is further complicated by joint sponsorship of data collection with non-government organizations that have objectives beyond those of exploration and discovery. This is a difficult issue since many exploration projects could involve funding from both federal and non-federal sources. In some cases, non-government sponsors provide their own data for merging with data collected with federal support. Forcing uncontrolled access to data whose collection was sponsored by non-federal participants would reduce the willingness of such groups to participate in federally sponsored OE program activities. These activities might also include government-sponsored researchers involved in traditional research activities where federal data management policy is less ambiguous. Expectations of data custody, ownership, and associated responsibilities for managing and distributing these data will be unique to each OE-sponsored mission and participant and must be clearly articulated in advance of the expedition and enforced afterward. The primary measure of ownership must be the source of sponsorship from which resources are used to create derived data. Again, in exercising its right of first use, NOAA and OE must recognize the consequence of public-private partnerships participating in OE program activities and related ownership and copyright issues. These data ownership issues must be clearly articulated in OE policy guidance and contract award vehicles intended for OE collaborators.

Federally funded investigators may own a copyright on the publication of processed data (i.e., Level 4 metadata) developed or bought under OE sponsorship. Any such publication would include a notice identifying the sponsorship and recognizing the license rights of the government under this clause. Pursuant to Circular A-110, NOAA reserves a royalty-free, nonexclusive, and irrevocable license to reproduce, publish, or otherwise use, and to authorize others to use, for federal government purposes, the copyright in any work or any

rights of copyright purchased by an investigator using federal funds. NOAA and OE retain the right to analyze, formulate, and publish summaries of ocean exploration data while allowing investigators the right to be credited for having collected and processed the data. In accordance with generally accepted academic courtesy standards, the investigator and the applicable collection platform should be acknowledged in subsequent publications that rely on any part of the data. Manuscripts resulting from OE-sponsored exploration activities that are produced for publication in open literature, including peer-reviewed scientific journals, should acknowledge OE sponsorship.

In cases where ocean exploration data or information could create risk or harm to public users from its loss, misuse, or unauthorized access or modification, Circular A-130 requires NOAA to protect these data or information commensurate with the level of risk and magnitude of harm that could result. This policy also requires NOAA to disseminate information to the public on equitable and timely terms while achieving the best balance between the goals of maximum usefulness of the information and minimum cost to the government and the public. Correct interpretation of this guidance could become critical in situations where a discovery made during OE program activities could provide large scientific or economic advantages to those who have access to the information, or could jeopardize national security. NOAA and OE guidance concerning responsibilities and actions to be taken by the OE program project coordinators in the event such a discovery is made must be clear and made available in advance of exploration activities.

In the context of federal data policy specific to the oceanographic community, NOAA Administrative Order 216-101, which provides standard guidance for oceanographic data, does not directly address ocean exploration data, the goal for its rapid and broad dissemination to support education and outreach, or retrospective requirements for managing ocean exploration data to facilitate follow-on research. As a result, the application of NOAA guidance to OE can be interpreted within the guidelines of more recently implemented federal policy. Under 216-101, federally sponsored investigators involved in OE program activities are obligated, and should be required by OE policy, to include within their proposals a description and itemized costs of data collection, processing—including quality

assurance, calibration, and format conversion,—and storage needs supporting their proposals. These itemized costs are critical since they become the foundation for data ownership and IPR decisions. OE must strike a balance between an investigator's need to hold data for analysis and to support publication of scientific results from these data and the fundamental goal of broad and timely public access to new discoveries. NOAA-sponsored investigators are required to provide public access to these data via a government approved archive facility within one year[35] of collection. Separate from this requirement, NOAA should encourage participating investigators, in the spirit of exploration and discovery, to provide copies of collected data for storage on the OE central repository and for archival as soon as possible following collection.

As discussed in Section 3.2.3, 216-101 requires NOAA-funded investigators to submit metadata for each data set as soon as possible after data collection. Metadata accompanying fundamental data sets made available for public access should include a disclaimer that the associated data are only as good as the quality control procedures applied by the PI and outlined within associated Level 3 metadata. An additional statement should note that users bear responsibility for the data's subsequent use or misuse in further analyses or comparisons, that the federal government does not assume any liability to users or third parties, and that the government will not indemnify users for liability due to any losses resulting from the use of the data.

OE program activities will inherently generate a variety of on-scene, primary and ancillary data that describe daily operations and activities, and involve direct support to education and outreach activities (e.g., still imagery, digital video and associated annotations, cruise logs, shipboard Science Computer System (SCS) logs on NOAA vessels, interview notes, etc.). These data include the majority of information necessary for generating a CSR and fulfilling Level 2 metadata needs. Generally, these data do not require significant processing, application of quality control techniques, or further detailed analysis by subject matter experts in order to be exploited for discovery and invention purposes. Except where pre-negotiated copyright agreements with partner organizations exist, these data should remain as real and intellectual property of NOAA and OE with copies available to investigators via

mission-specific agreements for data sharing. Imagery, video clips, notes, and logs should be made available to OE outreach and education components in as near-real time as possible using the OE public Web site as the primary vehicle. There may also be value in providing access to this subset of data via an OE MIS.

OE data policies must be an integral component of an OE Data Management Plan. OE must clearly state expectations of data custody, ownership, and responsibilities for managing and distributing ocean exploration data. These expectations must be tailored to each OE program participant since each contractual relationship will be unique. OE must also enforce established data management policies. The Data Management Plan should also include procedures for the collection, processing, and storing of exploration data and directions for access to data and products. Operational documents for each exploration activity, such as cruise instructions or deployment plans, should include detailed information on data management, collection, recording, and reporting responsibilities along with specific exploration objectives and the schedule of events.

## 3.5 Partnerships

There are a variety of government, public, private, domestic, and international organizations that are capable of conducting exploration activities, or that have a stake in the product of these activities. This fact, along with the multi-disciplinary nature of ocean exploration and the demand for a broad range of technological assets to support it, has resulted in multiple recommendations from independent sources for the creation of a national partnership for ocean exploration. This partnership would stand behind a shared set of strategies, goals, and responsibilities for the nation's ocean exploration program, advocate a shared plan above each member's self-interest, and ensure that information derived from exploration activities is accessible in the public domain.

The *Frontier Report* includes recommendations for establishing the management structure for a national ocean exploration program, including designation of a lead agency, using existing interagency mechanisms to ensure federal cooperation among agencies, and creating an Ocean Exploration Forum to promote public-private communication. NOAA, through its

OE program, has assumed the national leadership role in implementing these objectives and fostering collaboration among a broad cross-section of ocean exploration stakeholders. The existing charter of the National Oceanographic Partnership Program (NOPP) to integrate national efforts in ocean science and technology—including research and education—makes the NOPP an appropriate candidate for assisting NOAA in facilitating interagency cooperation that is beyond the scope of existing arrangements established by OE. NOPP initiatives related to ocean observations, standards, and oceanographic data management will help to ensure OE program compatibility with the national oceanographic research community. Recognizing these potential organizational relationships, NOAA and OE should continue to exercise their leadership role in ocean exploration and promote the broad spirit of exploration with NOPP through active participation in formulating processes and policy guidance as the national ocean exploration program evolves.

Within NOAA, there are many organizational elements with a stake in OE activities, including those within NESDIS, NOS, the National Marine Fisheries Service (NMFS), and NOAA Research. There are also many other government, non-government, public, and international partners and stakeholders that have the potential to contribute to or profit from OE program activities. Given the similarities between the concepts of exploration and the emphasis on outreach and education within OE and the NASA Oceanography Program, there will be many opportunities for collaboration and joint sponsorship of exploration initiatives and technology development projects of interest to both groups.

In addition to separately identifiable organizations, institutions, and commercial entities, the stakeholder community also includes the general public based on its associated educational needs (to increase involvement and proficiency in earth science disciplines) and the need for assurance of a positive return on its investment of tax dollars (to build a large public constituency that supports a vigorous national program of ocean exploration and research). Many non-governmental organizations with active outreach and education programs will find the OE program and its data management scheme a rich source of data and information.

There are a considerable number of initiatives within other government, non-government, commercial, and academic organizations with guiding principles that support an invigorated national emphasis on ocean exploration. Hundreds of organizations have been identified as having potential interests in data and information collected during ocean exploration. Many of these entities also represent likely partners in executing the OE program and potential participants in exploration activities.

While the principles and components of these many initiatives are too voluminous to describe in this document, they are important contributors to a national emphasis on ocean exploration. The OE will coordinate with this diverse community in its implementation of a data management strategy to maximize effectiveness and return on investment. These ocean exploration stakeholders and potential partners can be grouped into the following general categories:

- NOAA line offices
- National and international coordinating bodies (NOPP, UNESCO, CORE, UNOLS, etc.)
- Federal policy makers and legislative organizations
- Federal environmental science, energy, and medical research agencies
- Federal management, survey, and law enforcement entities (USGS, Minerals Management Service, U.S. Coast Guard, etc.)
- Military organizations (Office of Naval Research, U.S. Naval Oceanographic Office, Army Corps of Engineers, etc.)
- National historical societies (Smithsonian, National Geographic Society, etc.)
- National, regional, and local media
- Federal, state, and local marine sanctuary systems and protected areas
- Federal and commercial data repositories and data management centers serving research institutions and the public, including the Joint High Density Storage Association (HDSA) and National Institute of Standards and Technology (NIST) Data Preservation Test Facility[36]
- Professional research and technology societies (Institute of Electrical and Electronics Engineers (IEEE), Marine Technology Society, etc.)
- State and local governments
- State and local fisheries and marine resource management councils and organizations
- Oceanographic Institutions (Harbor Branch, Scripps, Woods Hole, etc.)
- Nonprofit oceanography, educational, and marine archeological societies, foundations, alliances, centers, laboratories, consortia, and associations

- Aquariums and marine museums
- Commercial energy, mining, fishing, and diving corporations
- Biotechnology corporations and commercial medical laboratories
- Commercial marine platform, sensor, and information technology development corporations
- Undergraduate and graduate oceanography and marine archeology programs
- K–12 education programs
- Maritime stewardship and conservation groups and unions
- Fine arts advocacy groups

## 3.6 Responsibilities

Achieving OE goals for data and information management requires that the roles and responsibilities of exploration partners and participants be articulated and aligned with OE objectives. For OE program activities, 216-101 requires all involved NOAA officials to be responsible for ensuring that the provisions of applicable data policy guidelines are followed.

As the line office within NOAA charged with operating the data centers and promoting critical environmental data and information services, NESDIS has responsibility as the long-term steward of OE data and information. NESDIS will need to enhance its facilities for archiving and serving these data in order to meet this responsibility. In moving towards compliance with national data policy, standards, and procedures, NESDIS leadership will be critical for strengthening partnerships and promoting wide public access to unique ocean exploration data. Planned enhancements for managing large amounts of data and information, such as the CLASS project, must be structured to accommodate emerging requirements resulting from OE data collection activities.

Because the various host exploration platforms have different equipment suites and standard operating procedures, a unique set of responsibilities will be generated for each individual ocean exploration expedition. These platforms are typically surface-borne exploration or research vessels operated by national organizations and academic institutions and organized under partnerships such as the University-National Oceanographic Laboratory System (UNOLS). For the purposes of data management, these vessels deploy and manage assets

that might be considered exploration platforms themselves, such as unattended buoys, submersibles, ROVs, AUVs, and acoustic arrays. OE has a responsibility to ensure that individual responsibilities for managing data are identified, tailored to each unique expedition, and clearly articulated in guidance documents and cruise plans.

Aboard these vessels, the Commanding Officer is responsible for the safe and efficient operation of the vessel and its assigned personnel, while the Chief Scientist (occasionally referred to as the Mission Chief or the Mission Coordinator) is responsible for the successful completion of exploration and research objectives outlined in the vessel's deployment plan. Additionally, the Chief Scientist or a designated on-scene data manager with responsibilities specifically assigned by OE directs the collection of Level 1 metadata during the conduct of an expedition. The value of a dedicated data manager working in conjunction with the Chief Scientist and participating PIs was recognized by the NOAA NOS during the FY01 Islands in the Stream expedition and led to a requirement for the inclusion of an *ad hoc* NOAA Data Manager within the complement of personnel participating in at-sea operations.[37]

The creation of an OE Data Management Implementation Plan based on this strategy will provide a means for establishing OE directives that govern the documentation, implementation, and assignment of data management responsibilities. It will also give OE project coordinators guidance to assist them in implementing exploration plans and supporting on-scene data disposition.

A data management function must be incorporated within the OE government staff to formulate and maintain data policy guidance and communicate regularly with OE program participants with data management responsibilities. This OE staff member will have oversight of the data management process from collection planning through archival, and will be an advocate for public and private users of exploration data. For example, this staff member would identify data that are not restricted by IPR considerations for immediate integration into the OE data management system. This rapid integration would facilitate quick access to these data by the public and support internal program assessment and measures of performance by OE.

# 4 ARCHITECTURE ALTERNATIVES

This section presents the architectural alternatives for OE data management. It identifies data management principles and associated architectural issues that must be addressed and incorporated into the OE program. It provides data management architecture alternatives, associated technical and data management issues and supporting technologies, and benefits, costs, and risks associated with each alternative.

Table 4-1 defines the platform nomenclature used in the supporting alternative architectures discussed throughout this section.

**Table 4-1. Platform Nomenclature for Alternative Architectures**

| Platform | Function | Data and Metadata Format |
|---|---|---|
| Exploration Platform Databases | Primary storage for data and metadata collected on an expedition. Includes ship log data. | Investigator standards supported by OE policy |
| Investigator Database | Data storage for exploration data and metadata that are in possession of investigator at the investigator's parent organization | Standards applicable to investigator's organization |
| OE Central Catalog | Searchable catalog containing Level 1, 2, and 3 metadata for all exploration data and products that were collected and developed under the OE program | Federal and NOAA standards (FGDC) |
| OE Central Repository | Data storage for recently collected, raw and derived exploration data that are in possession of NOAA | OE-approved standards |
| NOAA Archive | Long-term data archival for all exploration data, products, and associated metadata that are collected or developed within the OE program | NOAA standards |
| OE Portal | Primary web site for access to the OE catalog and entry point portal for accessing OE program data and products | OE-approved standards |
| OE Data User | User applications supporting user functions. Holds selected OE data downloaded from OE platforms | Standards applicable to user's organization |

Figure 4-1 illustrates a generic, high-level description of the movement of ocean exploration data as it relates to the functional processes of collection, processing, storage, access, and archiving. The dashed arrow connecting the NOAA archive to the OE Data User reflects the current, manually intensive process of accessing older data from the deep archive; this process is expected to improve as the benefits of NOAA programs—such as CLASS—are realized.

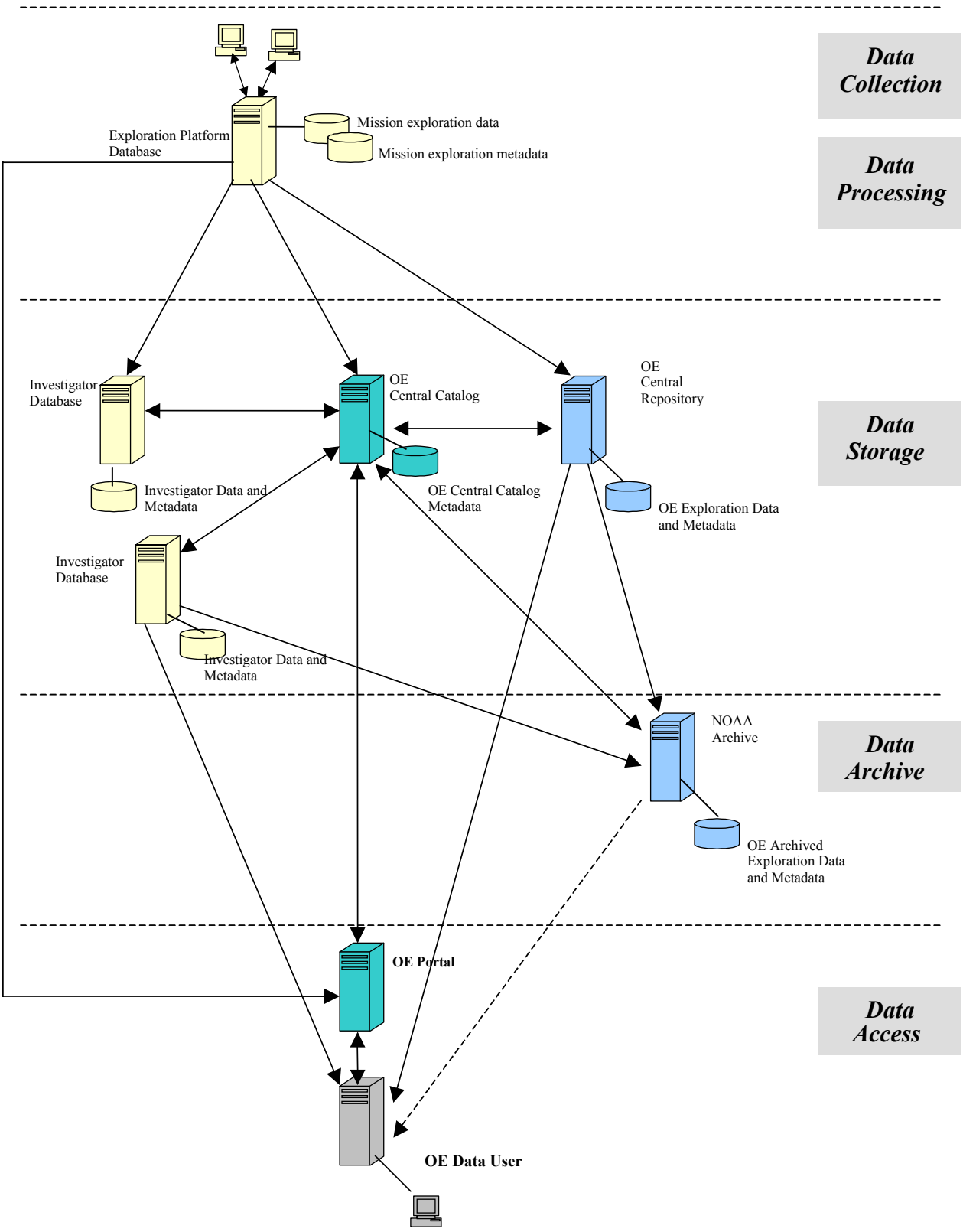



**Figure 4-1. Generic Movement of Ocean Exploration Data**

## 4.1 Architecture Principles

The architecture supporting an OE data management capability should use a set of supporting principles as developmental guidance. The following architecture principles are represented within this data management strategy.

- The OE program will produce data in a variety of digital and non-digital formats
- OE will manage a large volume of digital scientific data
- The public will be provided access to the OE data resident in a central repository
- The OE central catalog will include both data that is physically resident in a NOAA database and data that is virtually represented in this database but physically resident in databases at distributed locations
- Users of OE data will access large volumes of data over telecommunication links
- The OE program will provide access to a subset of non-digital data such as analog imagery, analog video, and biological and geological samples
- OE will catalog all data collected in conjunction with its sponsored activities, including data NOAA does not own and data stored outside the OE central repository
- OE will make its catalog available to other environmental data clearinghouses to maximize the utility of collected data
- OE will manage data stored in multiple, distributed locations
- Data resulting from OE program activities will be employed across multiple disciplines and in conjunction with data from other sources
- OE will permanently archive data as guided by NOAA policy using NOAA resources
- Level 1 and Level 3 metadata will be captured as data are collected and again at each subsequent stage of processing and storage
- Level 2 metadata (a CSR) will be generated after each expedition or other exploration field activity
- OE policies will govern submission of and updates to data and metadata resident in the central repository, catalog, and archives, as well as within the holdings of collaborating institutions

### 4.1.1 Data Accessibility

As discussed in Section 3.4, OMB Circular A-130 obligates OE to provide public access to data collected by the OE program. Public users must have remote access to the OE catalog and relevant data via the Internet or similar publicly available connection. It is also necessary that data be available to these users in standard formats and using standard protocols. For

most public applications, these formats and protocols should consist of commonly accepted commercial standards for on-line access to information, such as hypertext markup language (HTML), XML, Joint Photographic Experts Group (JPEG), and Motion-JPEG (MPEG). For scientific and other programmatic applications, data should be accessible on-line in widely recognized open standard formats and protocols that allow users to use a variety of compatible software manipulation tools. In particular, geographically referenced data—recognized in this strategy as comprising the bulk of ocean exploration data—should be made available using interfaces and protocols that are compatible with an appropriate spatial data model. This approach would allow complex information and services to be accessible and useful to a range of users employing many different types of software applications. The OE data management architecture should provide this open standards access using one or both of the following approaches:

- *Spatial Data Standardization*. OE data are stored in the central repository and in investigator databases using standard formats and protocols for spatial data
- *Spatial Data Translation*. The central repository provides a translation service using data exchange interfaces (DEIs) and a spatial data model accessible through the OE catalog, allowing users to access derived data in non-standard formats while accepting delivery in open standards formats

Under NOAA sponsorship, the NCDDC recently implemented an operational example of applying DEI capabilities for accessing distributed data in native formats.[38] By employing distributed object computing techniques coupled with a spatial data model in its middleware approach, the NCDDC provides access to distributed data stored in heterogeneous formats in different locations. This approach is consistent with the recommendations resulting from the 2001 MG&G workshop[39] and supports the recommendation for a transition to centrally managed, distributed, discipline-specific data centers that are developed, evaluated, and funded by cognizant government agencies. It also is consistent with national initiatives on spatial data infrastructures and the move toward open GIS frameworks to improve interoperability.

OE will manage data that is stored in multiple physical locations. Initially, data are likely to be in the PI's possession because of IPR considerations, or they may be turned over immediately for storage within the OE central repository. The data may continue to reside at

distributed locations for extended periods of time—even after access has been granted to the public—to support continuing research by users at those locations. Data may reside at specific locations and be organized by format, a likely scenario for video data if a NOAA-wide video library facility is established. Eventually, all raw data sets will be provided to the NOAA Data Centers for long-term archiving. OE policy guidance will define the responsibilities of PIs and organizations participating in OE-sponsored activities. This guidance may be tailored to each contractual relationship and will include the following elements:

- Direction to include data management costs necessary for satisfying OE requirements in investigator proposals
- Disposition and proposed location for data following the collection activity
- Recognition of any copyright or IPR considerations related to specific data
- Milestones for making data and metadata available to OE and the procedures to be used for submitting them
- Minimum acceptable performance standards for storage and telecommunications capabilities at distributed storage sites
- Data stewardship and backup requirements at distributed storage sites
- Identification of mandatory metadata elements and PI responsibilities for submitting metadata updates

The OE catalog will provide a centralized service, but its implementation will be physically distributed. Data accessible through the OE catalog may be distributed across NOAA Data Centers, program centers of data, research centers focused on specific disciplines, and the parent institutions of participating PIs. NOAA may not own certain data collected during ocean exploration activities since some organizations may claim IPR depending on levels of sponsorship and pre-negotiated collaboration agreements. Over time, the physical location of the data may change. For example, data may be relocated from the PI's collection equipment to a database at his or her parent organization to the OE central repository and eventually to an on-line, near-line, or off-line archive. The OE catalog contains Level 1 through Level 3 metadata, including the location of associated data at any point in time. This catalog indexes all data related to the OE program—including non-NOAA data—and can be used to track the data as soon as a metadata record is created, in some cases prior to data collection.

The catalog provides direct access to on-line data regardless of whether the data are held in the central repository or in an external collection. To provide this access, the catalog will be able to search compliant metadata that includes specifications for the physical storage location of the data and associated access mechanisms.

### 4.1.2 Telecommunications Capacity

Users will retrieve large volumes of data over telecommunication links. The OE data sets must be available to the public—as well as to federal, state, and local governments—to realize the benefits of the multidisciplinary discoveries that the program has potential to produce. This forum includes a variety of societies, museums, councils, laboratories, media groups, foundations, alliances, associations, and K-12, undergraduate, and graduate educations programs that require streamlined access to the data. Each user community will have a different level of interest in data collected under the OE program. For example, those involved in scientific research may have more interest in access to detailed raw and derived data, while interested media representatives and K-12 education programs may be more interested in products that represent derived data, imagery, and video.

While specific telecommunications design requirements are beyond the scope of this data management strategy, it is beneficial to address the general needs by separating the requirements of video from the remainder of the data. Telecommunications services at centers of data within NOAA are expected to be sufficient to satisfy the requirements for access to OE data. “Power users” of the data—those individuals using computer-intensive models and visualization techniques—will likely be physically collocated with their data at their home institution and will not routinely access their data through the OE central repository. Sophisticated content searches, sectoring, and compression at the OE central repository can further minimize impact on telecommunications requirements. Public users may experience loading delays as they download multiple video clips, such as MPEG files, containing specific content of interest. Expansion of telecommunications requirements for OE Web-based information beyond this level—such as integrating public access to raw digital video data—is not advisable because it would impact the broad cross-section of the public accessing information through the constrained bandwidth offered by dial-up capability. The provision of on-line access to fundamental video and imagery data is a requirement that extends outside of

OE and across all NOAA line offices. Developing a capability to address the storage, access, and archival demands of imagery and video is a specific recommendation within this strategy that must be addressed as a coordinated project within NOAA to ensure encapsulation of all requirements and identification of sufficient resources.

### 4.1.3 Cataloging Non-Digital Collections

The OE program will produce a variety of non-digital data, such as analog imagery and video, biological specimens, and geological samples. Cataloging of non-digital data is an important component that allows identification of physical data collected within the OE program, the current location of the physical data, and access procedures. The OE catalog will index non-digital (off-line) data and the associated metadata will include appropriate granularity to describe each individual item or group of items collected under the sponsorship of OE. Most of the information associated with analog imagery and video data, specimens, and physical samples is expected to evolve to a digital format.

### 4.1.4 Cataloging Non-NOAA Collections

The OE catalog will include data that NOAA does not own and data that is currently stored outside of the OE central repository. As a result, OE data policy must address data integrity issues at distributed storage locations to ensure accessibility and long-term stewardship. The OE catalog must be synchronized with modifications to data sets throughout the lifecycle of data. For example, when data stored at distributed sites are modified, lost, converted into a different format, or taken off-line, metadata maintained in the OE central catalog must reflect these events. The creation of additional derived data sets and descriptions of the additional processing applied to these data sets must be reflected by updates to associated Level 3 metadata. Data management policies and procedures must be developed that will support integrity of data across multiple locations. They must also provide a mechanism for maintaining and updating metadata and direct the level of involvement by PIs in this process.

## 4.2 Architecture Alternatives for Managing OE Data

This section describes alternatives for managing OE data within the context of the functional component themes illustrated in the data flow model (Figure 3-2 and Section 3.3.2 text) and prevalent throughout this strategy. Each functional component presents a set of specific data

management issues that can be resolved using alternative approaches. The OE selection of optimum alternatives are based on relative cost implications and a qualitative evaluation of the risks associated with each alternative.

### 4.2.1 Data Collection

Figure 4-2 illustrates the functional component of data collection. This component addresses the collection of data by PIs aboard vessels, other at-sea platforms, and shore-based exploration activities relying on remote sensors or sensor arrays.

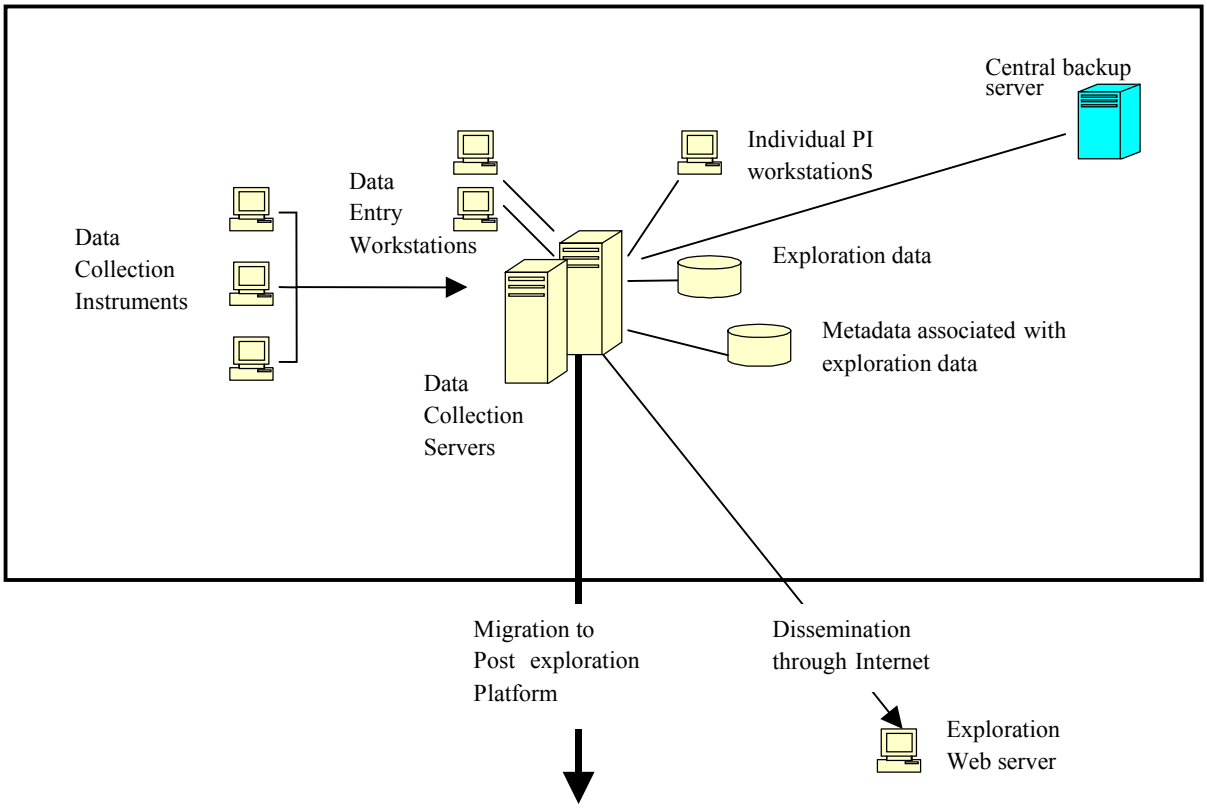



**Figure 4-2. Functional Component of Data Collection**

#### 4.2.1.1 Location of Data Collection Servers

Ocean exploration activities may be conducted in remote locations and may include multiple data collection platforms, such as vessel-based and hull mounted sensors, submersibles, ROVs, and AUVs. Data collected during an expedition may be stored on multiple native platforms, including PI workstations, workstations associated with individual instruments, a central coordinated server, or a coordinated server located at a geographically separate site. Data policies of the OE and a complimentary, standard concept of operations must address data

management issues related to prevention of data loss, data backup and recovery requirements, and enforcement of data standards. While a central server on a vessel would facilitate tighter data control and management, it is impractical to incorporate this approach into the majority of exploration activities because of the diversity of operations and associated technologies. A coordinated remote server for all collected data may not be practical due to the significant demand on telecommunications assets used to send and receive data between the exploration platform and the remote location. For specific applications, it may be appropriate and desirable to provide this link for a specific data type. One example might be a audio and video link between at-sea platforms and participants ashore, an arrangement that would allow real-time, multidisciplinary examination, analysis, and annotation of data by a large group of subject-matter experts who could not otherwise be accommodated on board the vessel. This type of link would also support direct Internet dissemination for education and outreach initiatives, such as Webcasts and real-time observation of ongoing activities.

#### 4.2.1.2 Creation of Metadata

An OE-identified minimum set of Level 1 through Level 3 metadata should be created at the time of data collection. The PI will be responsible for ensuring that required components of metadata describing data sets are created and provided to OE within the specified time constraints and format. Use of NOAA standard metadata creation tools, if available, should be encouraged. These tools should be able to generate labels that can be attached to physical samples and recording media for tracking purposes. Incorporation of metadata delivered by instruments and sensors as part of the data stream should also be encouraged. The collection of data to support the generation of a CSR (Level 2 metadata) will be the responsibility of the individual designated by OE to be accountable for managing this information, generally either the Chief Scientist or a specifically appointed on-scene data manager. This CSR will be delivered to OE within a specified timeframe. Attention to metadata standardization will facilitate migration of metadata to the OE catalog and permit uniform search capabilities for all OE-related data.

#### 4.2.1.3 Migration to Post-Exploration Platform

Following an expedition, data that has not already been transferred ashore via tailored telecommunications links will transition from its repository aboard the exploration platform

to a repository that is under the cognizance of the PI or directly to the OE central repository. OE policy and standard operating procedures will specify the procedures that guide this transition.

Table 4-2 identifies alternative approaches for resolving data management issues related to data collection. Preferable alternatives are indicated in bold type.

**Table 4-2. Alternative Approaches for Data Collection Issues**

| Issue | Alternatives |
|---|---|
| Location of Data Servers | Central on-site data server<br>Individual on-site data servers<br>Remote server |
| Creation of Metadata | **Standardized metadata collection**<br>Manual creation of metadata<br>**NOAA-sponsored metadata creation tools**<br>Automated interfaces to instruments |
| Migration to Post-Exploration Platforms | Platform-to-platform data transfer<br>Removable peripheral devices<br>Transportable storage media |

### 4.2.2 Data Processing

The data processing phase includes all PI activities that transform raw data into a format that has utility for follow-on use. These activities include processing of instrument data, error correction and quality control procedures, application of calibration data, protection and backup, and formatting and conversion. Although it is desirable to complete data processing requirements immediately after the collection process, in many cases the PI will require the additional time, assets, and computational resources available at host sites and discipline-specific research organizations following the mission. Table 4-3 identifies alternative approaches to resolve data management issues related to data processing. Preferable alternatives are indicated in bold type.

#### 4.2.2.1 Calibration

PIs will be required to perform any calibration procedures necessary to correct the raw data. OE should avoid storing and providing access to data that has not been calibrated since these data sets have the potential to misrepresent actual observed conditions. Calibration

information—such as applied procedures and coefficients—must be included within Level 3 metadata submitted by the PI.

**Table 4-3. Alternative Approaches for Data Processing Issues**

| Issue | Alternatives |
|---|---|
| Processing of Instrument Data | **Investigator processes data at collection time**<br>Investigator processes data post-expedition |
| Error Correction and Quality Control | **Correction of data occurs at collection time**<br>Correction of data occurs post-expedition<br>**Metadata updated with quality control procedures to data on-scene**<br>Metadata update occurs post-expedition |
| Calibration | **Correction of data occurs at collection time**<br>Correction of data occurs post-expedition<br>**Metadata updated with calibration data on-scene**<br>Metadata update occurs post-expedition |
| Protection and Backup | Central backup server and procedures<br>**Individual investigator backup procedures**<br>Backup to remote site |
| Data Formatting and Conversion | **Data formatted and converted at collection time**<br>Data formatted and converted post-expedition |

#### 4.2.2.2 Protection and Backup

To prevent loss of data, OE policy guidance should reflect standard procedures for data backup and recovery. The use of a central backup system would help to standardize the process but may be impractical because of the diversity of technologies and data formats. Additionally, implementation of backup procedures with all data resident on a single system and on the same platform would not prevent loss in the event of system failure or damage to the platform or vessel. PIs should be provided specific direction with regard to the protection of collected data. Backup procedures may include a periodic off-load of data to a remote, shore-based location for long-duration missions. This off-load could be accomplished via a telecommunications link or by physical delivery of backup media to the remote location.

#### 4.2.2.3 Formatting and Conversion

Data will be collected through the OE program from different data collection instruments that will produce data in various formats. The OE office should select and approve data and

metadata format standards that will be used across all OE programs and missions. Conversion to OE-approved standards should be done at the time of data collection.

### 4.2.3 Data Storage

This component supports data management after it has been collected and processed. These data are transitioned to an OE- or PI-sponsored system for further analysis. This phase is distinguished from archiving by the fact that data are placed into storage for the purpose of providing users direct, on-line access to the data. The length of time the data are maintained in storage is at the discretion of OE and depends upon the demand from the user community. Figure 4-3 illustrates the functional component of data storage. A specific objective of this strategy is to make OE data available for access by a broad cross section of public users within one fieldwork cycle (approximately one year) to ensure the utility of these exploration data is maximized.

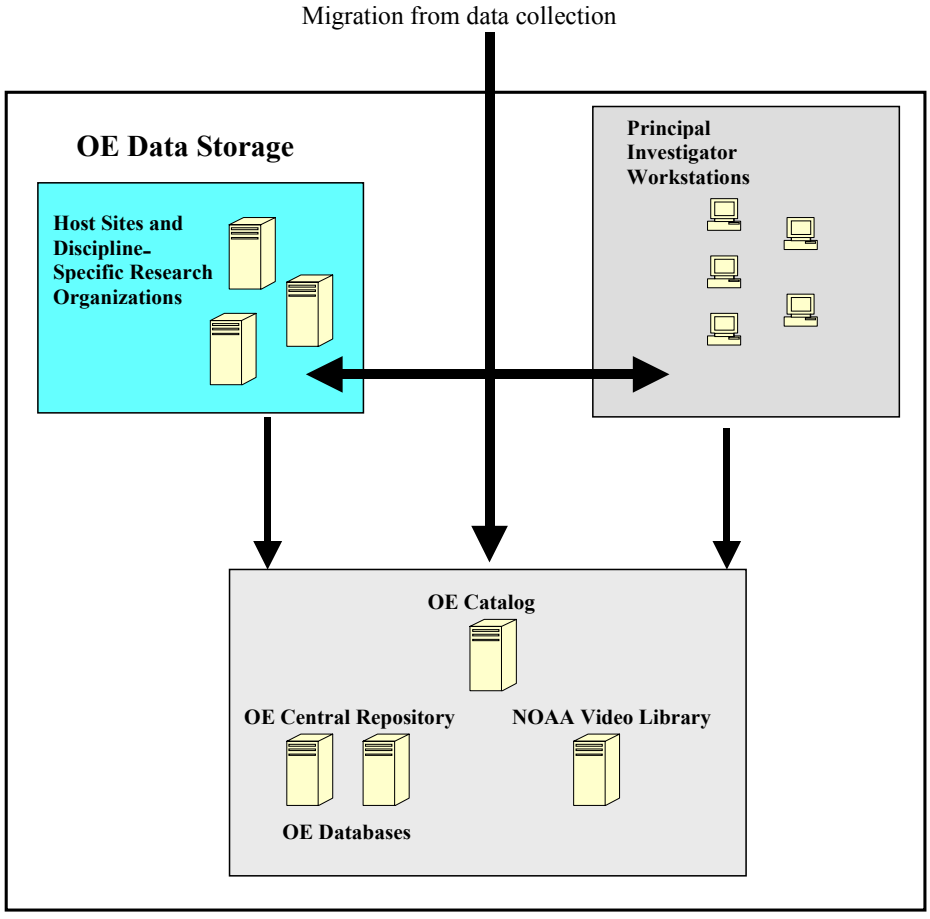



**Figure 4-3. Functional Component of Data Storage**

#### 4.2.3.1 Location of Data

There are multiple options for the storage of data in this phase. It is likely that a combination of options will be desirable. Specific responsibilities of PIs and other collaborators will be a component of OE policy guidance. The most likely options include the following:

- *All data are stored in the OE central repository*. While this option offers OE good control over the data and provides for quick archiving at the NOAA Data Centers, it does not provide for IPR considerations or facilitate a PI's stewardship of data and research activities during the post-collection period.
- *Data are stored on PI systems and delivered to OE and the NOAA archive after set periods of time*. This option recognizes an investigator's IPR to data and facilitates follow-on research. Since the data are out of direct OE control, policy guidance related to investigator responsibilities for maintaining the data must be explicit.
- *Data are stored in the OE central repository and replicated on PI systems*. This option helps ensure that data are preserved in their original form and also facilitates PI stewardship and research. It does not avoid the need to accommodate IPR considerations, doubles the required storage volume, raises data integrity and database synchronization issues, and risks a loss of control of data sets due to multiple copies.
- *Data are stored at investigator host sites or discipline-specific research organizations and delivered to OE and the NOAA archive after set periods of time.* This option is similar to data stored on investigator systems, although data maintenance responsibilities would fall on the host site or research organization rather than individual investigators. This may be an advantage if the site maintains a data management infrastructure that incorporates additional capacity and exploitation tools.

#### 4.2.3.2 Maintaining Data Integrity

OE will forfeit a level of direct control over management of data stored at a PI's host site or discipline-specific research organization. As a result, OE policy guidance as to the responsibilities of these sites related to OE data maintenance, usage, protection, and preservation must be developed and made available to all collaborators.

#### 4.2.3.3 Collocation of Data With Power Users

There will be a particular subset of OE data users—mostly within the science community—that will require significant manipulation of large volumes of OE data to support complex applications or numerical models. In these cases, collocation of data storage at the user's host location will be preferable to other storage options that would require remote access to the data.

#### 4.2.3.4 Data Compression

Implementation of data compression technologies is essential due to the large amount of data that will be collected through the OE program. Many of these data compression approaches are specific to individual data types. OE policy guidance should accommodate use of compression schemes for large volume data sets to control the costs of storing and exchanging data. To ensure the widest utility of the data, standard compression techniques—such as MPEG for digital video—should be encouraged.

#### 4.2.3.5 Imagery and Video Data

As discussed in Section 3.3.1.3, the OE program will generate a significant volume of imagery and video data. To support OE and other line offices producing increasing volumes of data, the NOAA data management infrastructure must be expanded to accommodate these data, as well as the processes of storing, classifying, processing, archiving, and providing users with innovative, functional access. NESDIS is currently investigating the requirements and resources necessary to establish a NOAA-wide video data management system (VDMS). This VDMS will be designed to allow the central management and archiving of imagery, video, and associated metadata at a dedicated facility, most likely under the cognizance of the NOAA Central Library. It will also provide access to the data to a broad range of users. This strategy includes a recommendation that OE support and participate in the development of this centralized NOAA capability due to the significant benefit to OE in managing these important exploration data and the added ability to apply advanced technologies to support education and outreach objectives.

#### 4.2.3.6 Backup and Recovery Procedures

OE policy guidance must include standard backup and recovery procedures that will prevent the loss of data in case of system failure. These procedures must prevent loss of data in all locations where OE data are stored. Implementation may require redundant data stores for data that is temporarily or permanently stored on databases at organizations outside of NOAA.

Table 4-4 identifies alternative approaches to resolve data management issues related to data storage. Preferred alternatives are indicated in bold type.

**Table 4-4. Alternative Approaches to Data Storage Issues**

| Issue | Alternatives |
|---|---|
| Location of Data | OE central repository<br>Principal investigator systems<br>OE central repository (on investigator systems during recognized IPR period)<br>OE central repository and replicated by investigators<br>**OE central repository (at host site or discipline-specific research organizations during recognized IPR period)** |
| Storage Formats | **Standardized metadata and data**<br>Distributed sites maintaining data in non-standard format |
| Data Integrity | **Policies for maintaining data integrity are implemented and enforced**<br>Procedures for maintaining data integrity are developed as needed |
| Collocation with Power Users | **Data are collocated with power users**<br>Data are located at a remote or centralized location |
| Data Compression | **Data compression approaches are coordinated by OE**<br>Data compression approach is at discretion of each location |
| Imagery and Video | **NOAA implementation of a VDMS**<br>OE implementation of a VDMS<br>Video and imagery data are managed without a VDMS |
| Backup and recovery | **Backup and recovery procedures are coordinated by OE**<br>Backup procedures are at the discretion of remote sites |

### 4.2.4 Data Access

The data access phase is of primary importance to the OE program since it directly supports the OE goal of reaching out to stakeholders in new ways. Figure 4-4 provides a depiction of the data access environment.

#### 4.2.4.1 Location of Data

Location of data will impact system performance and ease of access. Collocation of data with specific user communities will facilitate access to large amounts of data but will not support the multidisciplinary access necessary to take full advantage of data produced by the OE program. The concept and requirement for a distributed approach to data management was discussed in Section 2.3 and is a result of the variety and complexity of oceanographic data types that will be managed by OE.

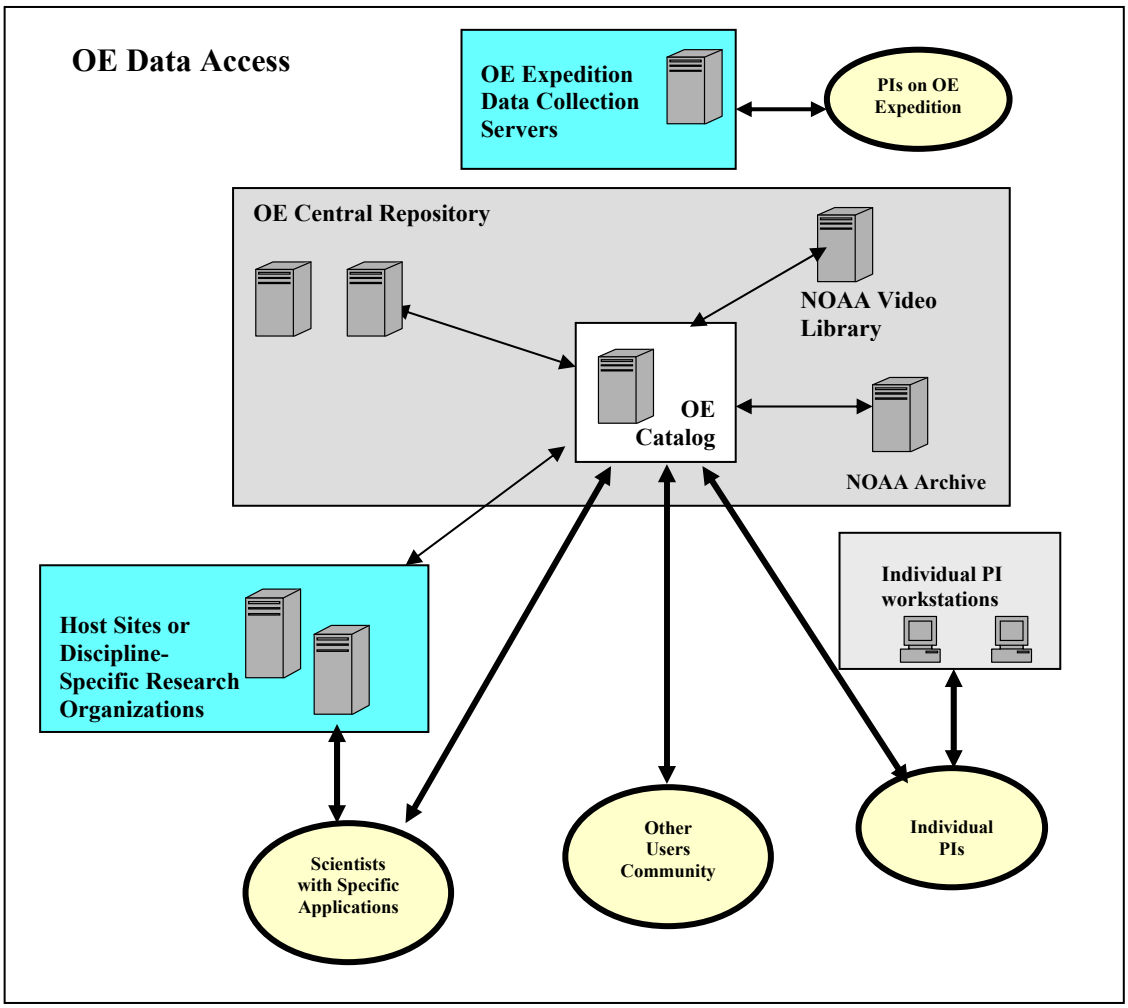



**Figure 4-4. Functional Component of Data Access**

#### 4.2.4.2 Metadata Format

In addition to complying with federal directives, metadata maintained in a standard format will enhance search, retrieval, and data exchange capabilities for all data users. OE should implement metadata standards discussed in this strategy for all data, including non-geospatial data.

#### 4.2.4.3 Maintaining Data Integrity

This strategy includes the storage and maintenance of OE data at distributed locations. OE policy guidance should include standard procedures for managing updates to metadata represented in the OE catalog for each modification to associated data sets and each additional derived data set that becomes available for access. In order to facilitate direct access to OE data

stored at remote locations, the OE catalog must be able to quickly reflect changes in data attributes and location.

#### 4.2.4.4 Access to Data by Different Classes of Users

As discussed in Section 3.5, a variety of users will seek access to OE data. In order to facilitate ease of access to a broad user base, the OE catalog should provide several different interfaces to accommodate different user classes. A scientist conducting basic research is likely to seek a depth of detail within the data and metadata that is not shared by the K-12 academic community. The OE portal should provide several options for access to data at different levels of detail. Much of the raw data collected during exploration activities will be used by PIs for specific applications purposes and may not be of significant interest to the public. Access to general information about ocean exploration activities and small subsets of derived data that are of high interest to the public should be represented on the OE public Web site without the requirement to access fundamental data via the OE portal at each use. Interface options should include the ability to focus a search on specific expeditions, timeframes, and research areas. Internet access with links to specific areas of interest would provide a straightforward interface for a variety of user communities. The OE portal should provide a central point of entry to data from all activities conducted within the OE program. Links to specific application areas—such as high-resolution bathymetry, new discoveries, and an annual atlas of OE accomplishments—should be included. These specific links would provide more detailed information and offer delivery format options for related data.

#### 4.2.4.5 Access to Data by Power Users

As introduced during the discussion of the storage function in Section 4.2.3.3, the subset of OE data users that require significant manipulation of large volumes of OE data to support complex applications or numerical models will also have unique access requirements. Since collocation of data storage at the user's host location is preferable to other storage options, it is likely that the local infrastructure will accommodate these power users with large-capacity local area networks or other direct access that can satisfy heavy access demands.

#### 4.2.4.6 Delivery Mechanism

An objective of this data management strategy is to provide the widest possible access to ocean exploration data via Internet or Internet-like capabilities. Due to the varied nature and volume of data associated with specific areas of interest, OE should also accommodate legacy, off-line delivery mechanisms such as distribution of compact optical disks. Table 4-5 identifies alternative approaches to resolve data management issues in the data access phase. Preferred alternatives are indicated in bold type.

**Table 4-5. Alternative Approaches for Data Access Phase Issues**

| Issue | Alternatives |
|---|---|
| Location of Data | OE central repository<br>**Distributed locations based on discipline-specific research and applications focus** |
| Metadata Formats | **Standardized format for all metadata**<br>Metadata format is determined by each location |
| Data Integrity | **OE policy guidance provides procedures for maintaining data integrity and accuracy of catalog**<br>Each site implements its own approach to maintain integrity between local data and the central catalog |
| Access to Data by Different User Classes | **OE catalog includes customized interfaces for different user classes**<br>OE catalog includes a standard interface for all users<br>OE provides access to data through focused individual web sites<br>**OE provides access to data via a central portal** |
| Access to Data by Power Users | **Data are collocated with power users for exploitation using local infrastructure**<br>Data are located at a centralized or remote location |
| Delivery Mechanism | **OE provides access to data on-line and accommodates legacy delivery media as required**<br>OE provides data access through delivery media used by data holder |

### 4.2.5 Data Archiving

As PIs make raw and selected derived data sets available to OE as required by established policy, OE will ensure that the data sets are submitted to the appropriate NOAA Data Center for long-term archiving. NOAA has designated NODC as the archive facility for physical, chemical, and biological oceanographic data and the NGDC as the archive facility for geophysical, geological, and geochemical data. Consistent with the recommendations of the 2001 MG&G workshop, all raw data sets will be archived. Companion data sets that have

been derived through the application of processing, quality control, calibration, or conversion techniques will also be provided to the appropriate archive facility. OE will comply with established NODC and NGDC policies and procedures for data archiving. Figure 4-5 illustrates the data archiving phase.

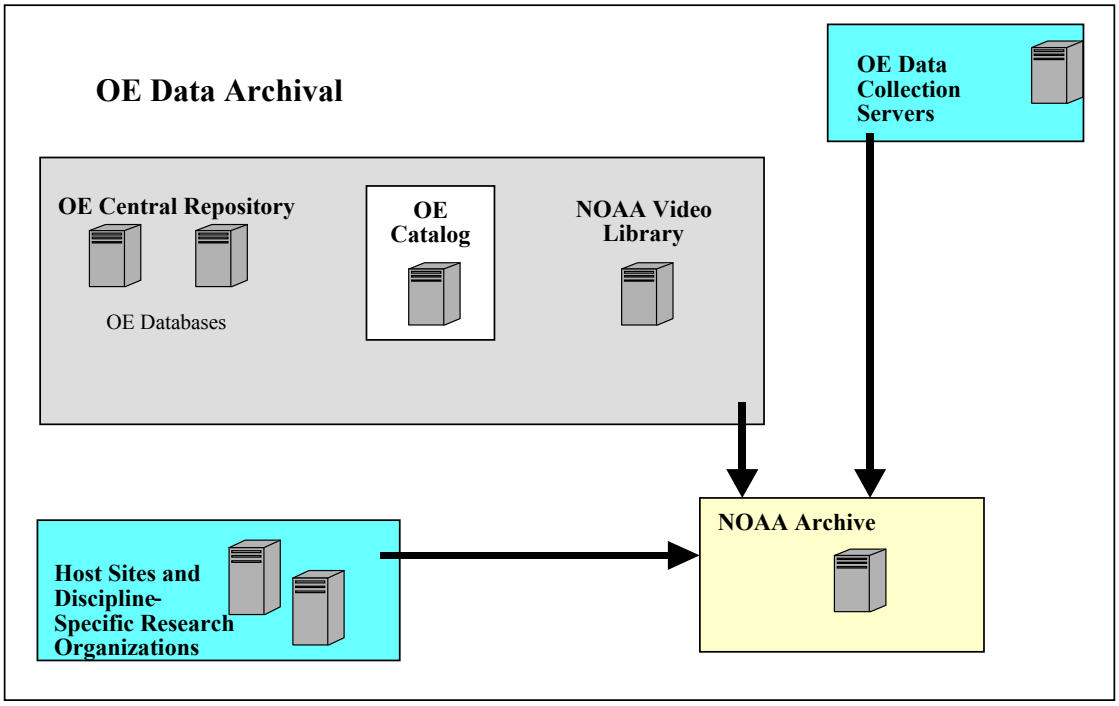



**Figure 4-5. Functional Component of Data Archiving**

#### 4.2.5.1 Large Volume of Data

The OE program will collect large volumes of data. NOAA is responsible for the long-term stewardship of the data and exercises this responsibility through its NOAA Data Centers. OE needs to work closely with NODC and NGDC to ensure that these archive facilities anticipate future data volumes and types that will be arriving for archival. Additionally, OE should seek an approach and procedures that would preserve non-NOAA data that directly support OE goals and objectives.

#### 4.2.5.2 Data Compression

Due to the large amounts of data that will be collected through the OE program and by all NOAA line offices, it is likely that NESDIS will need to incorporate data compression

technologies within NODC and NGDC. Data compression techniques implemented by NODC and NGDC will apply equally to OE data.

#### 4.2.5.3 Changing Technologies and Data Formats

NOAA will continue to improve the capabilities of its archive facilities as allowed by the level of technology and resources available. Initiatives such as CLASS are designed to allow NOAA to keep pace with the rapidly growing volume of data and the desire to improve user access to the archives. As new technology improves user access to NODC and NGDC archives, OE may need to consider modifying its data management strategy to reduce user dependence on the OE central repository and increase access to the NOAA Data Centers. Based on the current rate of change in applicable technologies and budget realities, the strategy may not need to be modified for a decade or more.

### 4.2.6 Alternative Architectures

Analysis of the individual functional components presented in the preceding sections leads to the following candidate alternative OE data management architectures: an OE central repository and catalog, an OE distributed repository with a centralized OE catalog, and an OE central repository with replication of data at host sites and discipline-specific research organizations. These alternatives are described in the following sections.

#### 4.2.6.1 Alternative 1: OE Central Repository and Catalog

Figure 4-6 illustrates this alternative and the resultant movement of data. In Alternative 1, all data collected through the OE program are maintained in a central OE repository maintained by OE. At the conclusion of exploration activity, the data are stored on individual PI systems, while associated metadata are made available to the OE catalog. Pursuant to OE policy, when the conclusion of the recognized IPR period is reached, investigators forward data and metadata modifications to OE for storage on the OE central repository and for archiving in NODC or NGDC as appropriate. The government maintains its rights to these data throughout all phases; the PIs have temporary stewardship and IPR to the data during their research. OE maintains a central catalog that serves as a single point of entry for users of exploration data.

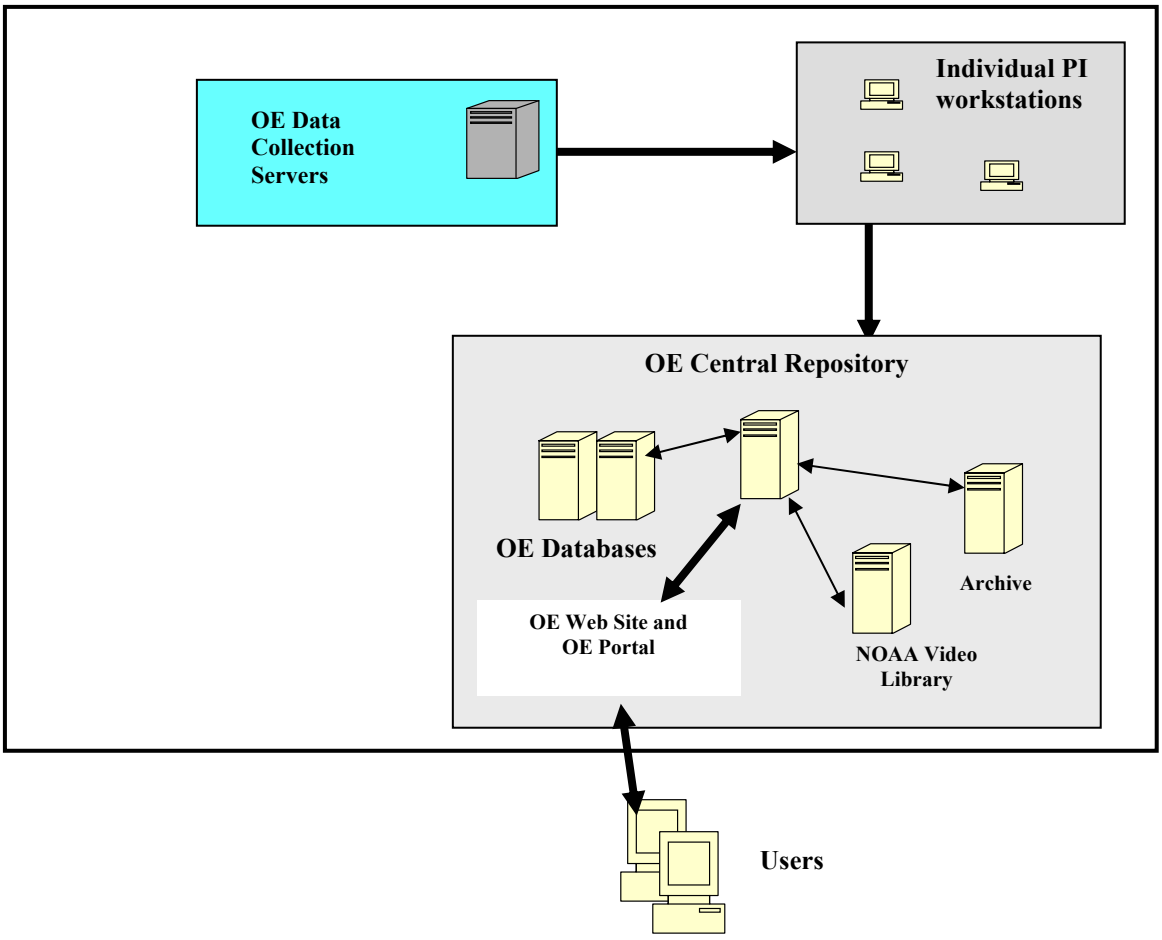
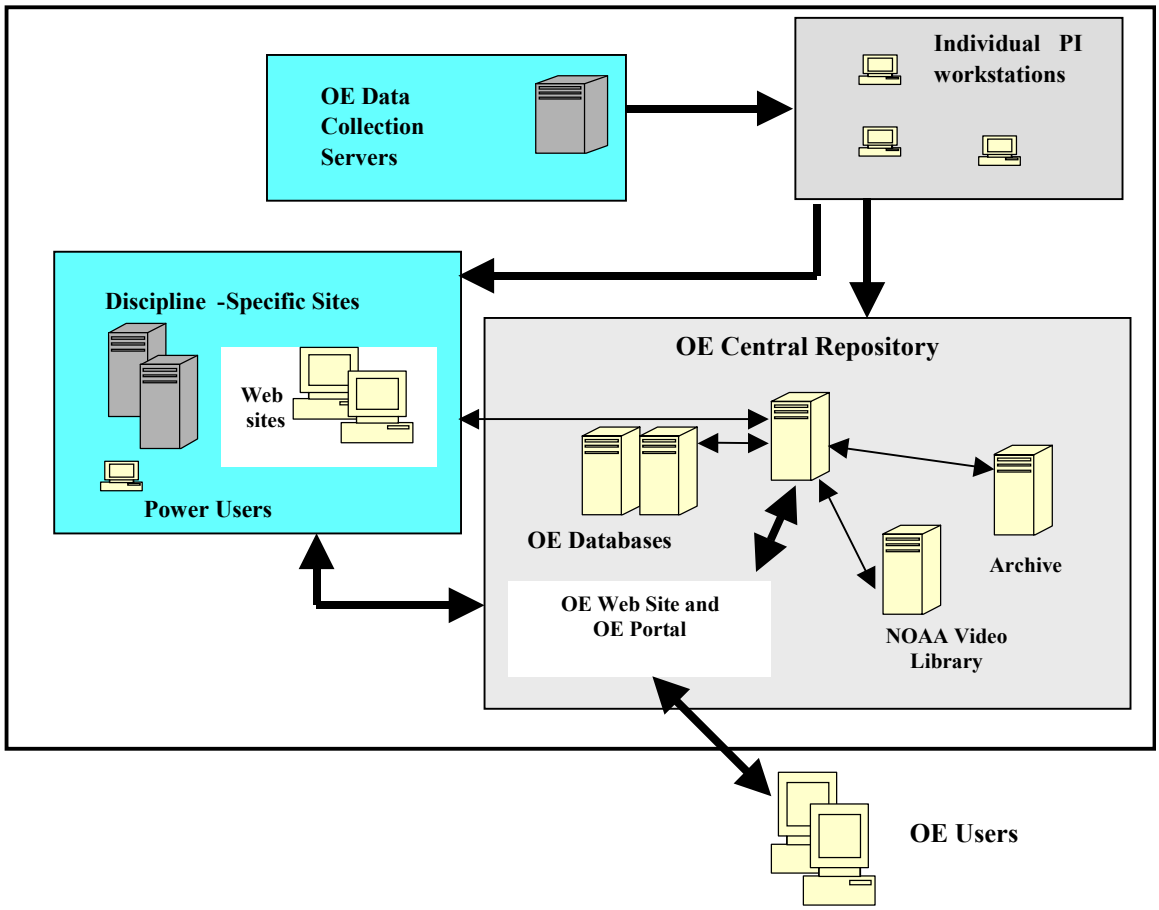



**Figure 4-6. Alternative 1**

#### 4.2.6.2 Alternative 2: OE Distributed Repository with a Centralized OE Catalog

Figure 4-7 illustrates this architecture alternative and the associated movement of data. In this alternative, PIs and collaborators at host sites and discipline-specific research organizations maintain all data collected through the OE program that is not directly forwarded to the OE central repository. These locations are based on their relationship to the investigator and the site's alignment with the applicable research focus. The OE central repository maintains data that are not aligned with any specific research entity. At the conclusion of an exploration activity, the data are stored on individual PI systems, while associated metadata are made available to the OE catalog. During the period when data are stored at distributed sites, the responsible manager at each site provides updates to the OE catalog as required and may provide direct access to data at any time during its residence. Pursuant to OE policy, when

the recognized IPR period is over, investigators may forward data and metadata modifications to OE for storage on the OE central repository and for archiving in NODC or NGDC as appropriate. Alternately, the host site may continue to provide on-line access to the data at the local site via the OE catalog and fulfill archiving responsibilities by forwarding copies of the original data to NODC or NGDC. The government maintains its rights to these data throughout all phases; the PIs have temporary stewardship and IPR to the data during their research. OE maintains a central catalog that serves as a single point of entry for users



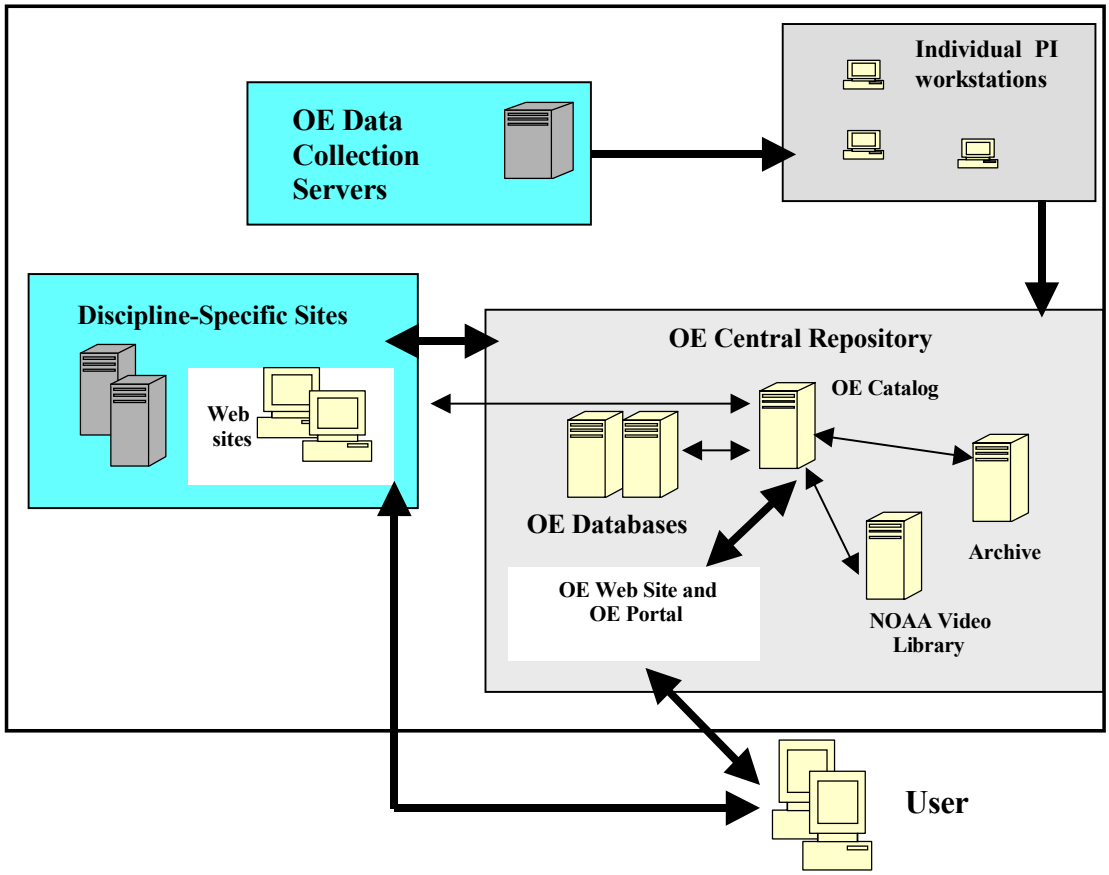

of exploration data.

**Figure 4-7. Alternative 2**

#### 4.2.6.3 Alternative 3: OE Central Repository and Catalog with Replication of Data at Host Sites and Discipline-Specific Research Organizations

Figure 4-8 provides an illustration of this alternative. All data collected through the OE program in Alternative 3 are maintained by OE in an OE central repository. It is also

replicated at host sites and discipline-specific research centers based on the site's relationship to participating PIs and the research focus of each site. At the conclusion of an exploration activity, the data are stored on individual PI systems, while associated metadata are made available to the OE catalog. Data are stored in the OE central repository after collection and replicated at discipline-specific sites as necessary. Pursuant to OE policy, when the conclusion of the recognized IPR period is reached, investigators may forward data and metadata modifications to OE for storage on the OE central repository and for archiving in NODC or NGDC as appropriate. The government maintains its rights to the data throughout all phases; the PIs have temporary stewardship and IPR to the data during their research. OE maintains a central catalog that provides a capability to search metadata and access data maintained on the




OE central repository.

**Figure 4-8. Alternative 3**

### 4.2.7 Assessment of Alternative Architectures

Table 4-6 provides a qualitative assessment of the alternative architectures. Each criterion is discussed relative to the alternative architectures.

**Table 4-6. Qualitative Assessment of Alternative Architectures**

| Assessment Criteria | Alternative 1 | Alternative 2 | Alternative 3 |
|---|---|---|---|
| NOAA Storage Requirements | - | + | - |
| NOAA Telecommunications Requirements | - | + | 0 |
| OE Control of Data | + | - | 0 |
| Data Integrity Management | + | 0 | - |
| Data Security and Preservation | + | - | 0 |
| Administration and Coordination Effort | + | 0 | 0 |
| Data Availability | - | 0 | + |
| Ease of Access | 0 | + | + |
| Access for Power Users | - | + | + |
| Cost to NOAA | - | + | - |

**Legend:** + alternative provides a good solution for this area
- alternative will have a negative impact on this area
0 alternative is neutral in this area

- *NOAA Storage Requirements*. All data are stored on NOAA platforms with Alternatives 1 and 3. NOAA stores only data that is not maintained by other host sites and discipline-specific research organizations with Alternative 2. NOAA will archive all OE data with all alternatives. Alternative 1 and 3 will have a negative impact on NOAA through increased demand on the NOAA infrastructure. Alternative 2 provides the best solution with minimum impact on the NOAA infrastructure.

- *NOAA Telecommunication Requirements*. NOAA handles all telecommunication loading for delivering data to the end user with Alternative 1. Most users from the research community access data directly through a location outside of NOAA with Alternative 2. Data can be accessed through NOAA or directly through locations outside NOAA with Alternative 3; power users from the research community will access data directly through a location outside of NOAA. Alternative 2 provides the best solution with minimum impact on NOAA telecommunication infrastructure. Alternative 1 will have a negative impact (increased demand) on NOAA telecommunication infrastructure.

- *OE Control of Data.* NOAA has full control over OE data with Alternative 1, as all data will be stored on the NOAA platforms. NOAA has only partial control over data with Alternative 2, as data are stored outside NOAA. Data are replicated on the NOAA site and locations outside NOAA with Alternative 3. NOAA lacks some control over data that is stored on locations outside of NOAA with this alternative. Alternative 1 provides the best solution. Alternative 2 will have a negative impact as NOAA does not have full control over data that is stored outside NOAA.

- *Data-Integrity Management*. Alternative 1 provides the best environment for data-integrity management since all data are stored at NOAA sites. Alternative 2 presents a challenge to data-integrity management since data are stored on locations outside NOAA, and NOAA can only recommend and issue guidance for maintaining integrity of data. NOAA has full control over a replica of data stored at the NOAA sites with Alternative 3; however, this alternative increases data-integrity maintenance effort since data has to be synchronized between remote locations and NOAA.

- *Data Security and Preservation*. NOAA has full control over data, backup and recovery procedures, and access control to data with Alternative 1. NOAA does not have control over data and cannot guarantee security of data or prevent loss of data with Alternative 2. NOAA can manage security of data that is stored at the NOAA sites but cannot guarantee the security of data that is stored at locations outside of NOAA with Alternative 3. Alternative 1 provides the best solution for data security. Alternative 2 will increase the OE data security management effort.

- *Administration and Coordination Effort*. Alternative 1 will require the least coordination and administration effort (e.g., enforcement of standards or maintenance of multiple versions of the same data set). Alternatives 2 and 3 will require more coordination effort as data are stored and maintained at multiple locations.

- *Data Availability*. NOAA can guarantee access to data since all data are stored on the NOAA sites with Alternative 1; however, most OE users will not be collocated at the NOAA sites. Availability of data will be at the discretion of sites outside NOAA with Alternative 2, but this alternative will facilitate access to OE data for specific user communities collocated with data. NOAA can guarantee access to a replica of the data that is maintained at the NOAA site with Alternative 3, and remote sites will facilitate access to OE data for local users. Alternative 3 provides the best approach for making the OE data available. Alternative 1 may have a negative impact on data availability due to remote location of the OE users.

- *Ease of Access*. NOAA is in control of data and can provide a standardized interface to all users with Alternative 1. NOAA can provide a standardized interface through the central catalog with Alternatives 2 and 3 but cannot mandate a standardized interface for remote sites. Alternatives 2 and 3 allow remote sites to provide customized interfaces that are best suited for a specific user community.

- *Access for Power Users*. Some research communities have high demand for the large amounts of data collected by OE. In those cases, collocation of data with a research community will benefit power users. Alternative 1 does not provide that opportunity since all data are stored at NOAA sites. Alternatives 2 and 3 provide the best solution through local access to OE data for power users.

- *Cost to NOAA*. While a cost analysis during the systems engineering phase would reveal more specific costs, this area simply compares the relative cost implications between alternatives. Alternative 1 would be most costly to NOAA as all data are stored at, and accessed through, NOAA sites. It would require additional storage devices, potentially

new servers, and would put an additional load on the telecommunication infrastructure and NOAA employees. The load and operational cost is shared between the NOAA and sites outside of NOAA with Alternative 2 and to a lesser degree with Alternative 3.

### 4.2.8 Risk Assessment

This section provides an assessment of certain risk factors that should be considered in selecting an architectural alternative for managing the OE data. Table 4-7 provides a summary of these risk factors.

**Table 4-7. Summary of Risk Factors**

| Risk area | Alternative 1 | Alternative 2 | Alternative 3 |
|---|---|---|---|
| Loss of data | Low | Medium | Low |
| Loss of data integrity between data sets and central catalog | Low | Medium | Medium |
| Loss of access to data | High | Low | Low |
| Performance degradation | High | Low | Low |

**Legend:** **High** - an unacceptable level of risk; must be mitigated or alternative is discounted
**Medium** - a significant level of risk; mitigation measures should be actively pursued
**Low** - an acceptable level of risk; mitigation measures may be pursued as desired

- *Loss of Data*. NOAA has responsibility for OE data preservation. Data collected through the OE program can be lost as a result of human errors, equipment failure, or improper data management procedures (e.g., backup and recovery procedures, procedures for returning data to NOAA from temporary custody by PIs, or data archiving procedures). Each alternative architecture provides some degree of risk for data loss. Alternative 3 provides the best solution for mitigating a data-loss risk through maintaining a replica of all data between NOAA sites and remote locations. OE data are under full control of NOAA with Alternative 1, stored on the NOAA sites. This facilitates implementation and enforcement of data management procedures for OE data. NOAA does not have full control over all OE data with Alternative 2 since data are stored outside NOAA and implementation and enforcement of data management procedures are under the control of the individual research centers. However, NOAA can mitigate this risk through active policy oversight and enforcement, and archiving OE data at the NOAA archives as soon as data are returned by individual PIs.

- *Loss of Data Integrity Between Data Sets and the Central Catalog*. The OE central catalog maintains a searchable subset of metadata with pointers to individual data sets. If data sets have been changed or relocated, these events must be reflected in a central catalog. Maintenance of data integrity between individual data sets and the OE central catalog requires development, implementation, on-site maintenance, and enforcement of data management procedures for updating the catalog. Implementation and enforcement of these procedures is simplified when the data and the OE central catalog are under full

control of NOAA. Alternative 1 is the best solution for mitigating this risk since NOAA maintains the OE central catalog and all data sets. Alternatives 2 and 3 will require additional coordination effort to minimize this risk.

- *Loss of Access to Data*. Temporary outages in the communication infrastructure or temporary outages in the supporting data servers can cause loss of access to OE data. Alternatives that provide different access paths or provide for the distribution of data on multiple sites provide the best approach to mitigating this risk. Alternative 3 provides the best solution to mitigate this risk since data are replicated at NOAA sites and individual research centers. Alternative 2 mitigates this risk by distributing data to multiple sites and locating the data close to the user community. Alternative 1 represents the highest risk due to its centralized dependencies, and would require mitigation through mirror sites or by providing alternate sites with selective replication of data.

- *Performance Degradation*. The OE user community represents a variety of users with diverse interests. An accurate estimate of the overall demand for OE data at this early stage of the program is difficult. Due to character of data (significant amount of graphical data, images, videos, and potential data products), access to data by the user community may present a significant load on supporting infrastructure and could result in performance degradation. Alternatives that provide data distribution over multiple sites are the best in mitigating the risk of performance degradation. Therefore, Alternatives 2 and 3 incur less performance degradation risk.

## 4.3 Recommended Alternative

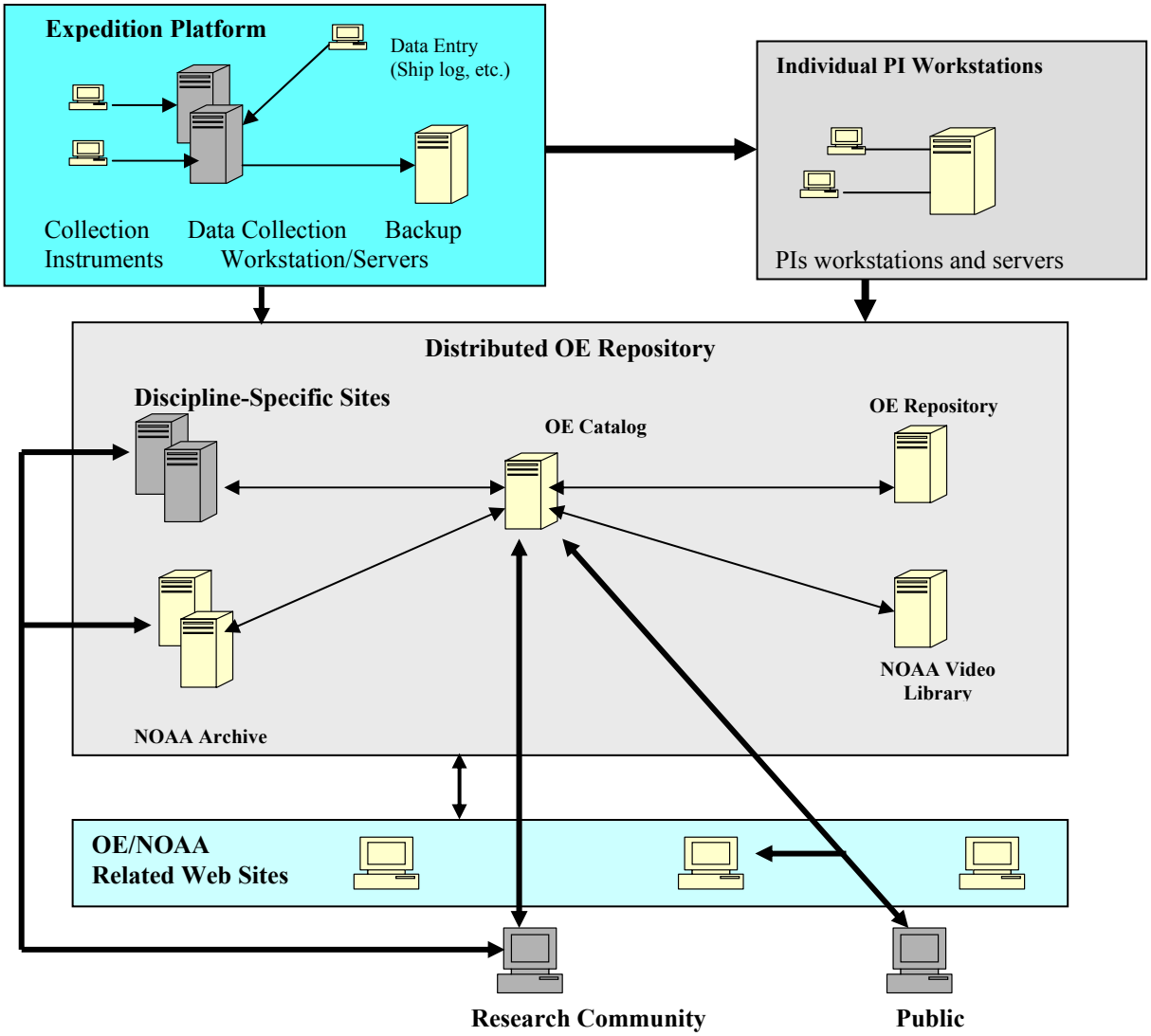
The architecture alternative based on a distributed OE repository with a centralized OE catalog is the most appropriate for supporting the program. With this alternative, data are stored in various host sites and discipline-specific research organizations and the NOAA Data Centers. NOAA will maintain a central catalog that contains a searchable subset of metadata pointing to a location of individual data sets, and provide a single entry point and access mechanism for data users. To preserve the data over time, NOAA will also provide an archiving function for collected data.

### 4.3.1 High-Level Architectural Design

Figure 4-9 provides an illustration of the high-level architectural design for the recommended alternative. The components of this design are briefly described in this section.

### 4.3.2 Architectural Components

The components of the recommended alternative architecture illustrated in Figure 4-9 support the phases of OE data management from collection through archival.




- *Collection Instruments*. Collection instruments are mostly commercial off-the-shelf (COTS) equipment delivering data in digital format according to manufacturer specification.

**Figure 4-9. High-Level Architectural Design**

- *Data Collection Workstations and Servers*. Data collection workstations and servers are mostly COTS components that will store and maintain data and associated metadata collected throughout an exploration activity. Individual PIs participating in OE expeditions may provide data collection servers as separate workstations or as components of systems accompanying collection instruments. NOAA will likely provide data collection servers for exploration conducted by NOAA personnel and for data collected from ship logs. A typical configuration for a PI might consist of common desktop systems and operating systems with sufficient peripheral storage capacity to store and manipulate data collected during the exploration activity.

- *Collection Backup Server*. A collection backup server is an optional component provided either by OE, the exploration platform, or by the PI for protecting data collected during an exploration activity. This component may also serve as a backup processor in case of failure to the original equipment. A typical configuration would consist of a common desktop system and operating system with sufficient peripheral storage capacity to store data collected from an exploration activity.

- *Data Entry Workstations*. Data entry workstations may be provided by PIs participating in OE expeditions to collect the data and metadata, in addition to the capabilities offered by their collection instruments. NOAA may also provide workstations for collecting and organizing data and metadata during the conduct of the expedition.

- *Principal Investigator Workstations and Servers*. The primary function of the PI workstations and servers is to store data that is in the PI custody immediately following collection activities and during individual research periods. PI workstations and servers will be provided by individual PIs and associated organizations. As with discipline-specific sites, the OE central repository, and the NOAA archive and video library, PI workstations and servers may provide direct access to local data via the OE central catalog. In Figure 4-9, these workstations and servers are illustrated separately since there are no restrictions on their location and accessibility to the data hosted on these assets is not guaranteed. In many cases, PIs will choose to host data on servers within the information technology infrastructure at the local site or discipline-specific research organization rather than the PI's server due to their additional capabilities and capacity to manage these data.

- *OE Central Catalog*. The primary functions of the OE catalog are to maintain metadata for all data collected under the OE program, to maintain links and access paths to data locations, to maintain a search capability based on multiple selection criteria, and to maintain an access or delivery mechanism for data that are of interest to the user community. A configuration of the OE catalog server will be determined based on the user demand. A typical configuration would be a group server running a common desktop operating system with sufficient peripheral storage capacity to maintain the OE catalog database along with the security to maintain the integrity of the catalog. The configuration should include an alternate server to minimize down time and to balance the load.

- *Discipline-Specific Site Servers*. The primary role of these servers is to maintain OE data for a specific OE outreach, education, technology development, or research focus. Configuration of these servers will be determined by individual host locations. These servers will have direct connections to the catalog using specified paths and in accordance with appropriate security measures. Use of direct connections and local infrastructure will accommodate power users at the host site with specialized, demanding data access needs.

- *OE Central Repository*. The central repository supporting OE will maintain the OE data that are not maintained by one of the discipline-specific sites. This includes data products associated with specific exploration activities, annual atlases of OE accomplishments,

and data provided to OE by PIs early in the management cycle while operational access—as opposed to archival—is still desired. Configuration of these servers will be determined based on amount of data collected and the products intended to be developed. The OE central catalog will provide direct access to data in this repository.

- *NOAA Archive*. The primary role of the NOAA archive—represented by NODC and NGDC—is to archive all data collected through OE programs and to preserve data for future generations.

- *NOAA Video Library*. The primary roles of the NOAA video library are to maintain and provide access to video data and associated metadata collected by all NOAA line offices (including OE), to provide a search function using multiple selection criteria, and to provide access to NOAA video data by the public.

- *OE Web Site and Related Web Sites*. NOAA and the discipline-specific research organizations may provide multiple Web sites focused on specific OE expedition areas of interest. Configuration of these services will be based on the type of information, the amount of data, the category of data (e.g. video, graphics), and the anticipated user interest.

- *Research Community*. The exploration research community comprises a variety of stakeholders from academia and other public and private laboratories and institutions with an interest in OE data. These stakeholders employ systems that allow direct access to OE data residing in distributed repositories and bypass the OE central catalog. These direct connections and local data management infrastructures can support the specialized data access needs of the research community. Connections to discipline-specific sites and the NOAA archive as a separate data access path is illustrated Figure 4-9 and highlights access paths that can be accommodated by tailored data handling systems in use within the oceanographic research community. An example of such a system in wide use by the oceanography community is the Distributed Oceanographic Data System (DODS).[40] DODS has the capacity to link data-handling applications with data sets in distributed locations. With specific linkages established by participating data users in the research community, DODS could also serve OE data.

### 4.3.3 Concept of Operations

When complete, the OE data management concept of operations will describe multiple phases of the OE functional process, roles and responsibilities in this process, operational procedures and policies that will be followed, and supporting infrastructure that will support OE in each phase of the functional process.

#### 4.3.3.1 Data Collection

OE data will be collected through OE expeditions. The data will be collected from multiple sources and in a wide variety of formats. For example, data can be collected through

collection instruments, observations and follow-up data entry, videos and images, space-based remote sensors, ship and event logs with information about each event during the OE expedition, research notes by OE expedition participants, and physical samples. In addition to raw data, each data and physical sample set must be accompanied by an appropriate set of Level 1 metadata (e.g., location, time, collection method, etc.). As much as possible, data should be captured in digital form at the time of the collection event. Data will be stored on data collection servers that may include systems provided by the PI. Data will be secured against data loss through appropriate backup and recovery procedures. At the end of an OE expedition, PIs will typically assume temporary custody of data, for a time period to be stipulated by OE policy guidance, to allow for organizing, processing, application of quality control procedures, and to support the PI's individual research. NOAA may assume custody of data collected by NOAA investigators, ship logs on government vessels, and copies of video and images that were collected during the expedition. Copies of video and images may also be provided to an individual PI if these data represent a unique supporting source of data for their research. Raw and derived data and information must be provided to NOAA at the end of the designated PI work period stipulated in OE policy guidance and in the PI's contract. Participants will be encouraged to submit data to NOAA as soon as is feasible in order to maximize the public good of sharing data.

- Design consideration: Provide NOAA-sponsored metadata tools on the research vessels
- Design consideration: Integrate functionality of metadata collection with ship logs and facilitate the development and installation of shipboard integrated measurement systems to collect shipboard data and metadata including data from "flow-through" systems
- Roles and responsibilities
  - Expedition Chief Scientists and PIs are responsible for data and metadata collection as directed by OE
  - The Chief Scientist, assisted by the vessel commanding officer, will be responsible for maintaining ship logs as directed by the OE cruise plan
  - OE will be responsible for providing support for data backups and recovery during expedition activities
- Operational policies and procedures that need to be developed
  - Agreements on IPR, data ownership, and temporary custody of data by PIs
  - Identification of the minimum set of metadata that must be developed by the PI and directions for submission
  - Procedures for maintaining required ship and data logs
  - Directions for producing and submitting a CSR

- Procedures for preventing data loss in case in equipment failures

#### 4.3.3.2 Data Storage

At the end of the OE expedition, PIs will take temporary custody of OE data to complete their organizing, processing, application of quality control procedures, generation of Level 3 metadata, and to conduct their research. While PIs may have temporary custody of OE data up to one year, they will have an obligation to deliver metadata to OE within 60 days following the conclusion of the expedition. OE will use this metadata to populate the central catalog. As individual PIs complete their exploitation of data under their cognizance, they will forward raw and derived data back to OE. Depending on type of data and the available transfer path, data will be stored at the PI's host site, a discipline-specific research organization, in the OE central repository, or in one of the NOAA Data Centers. The OE central catalog will be updated to reflect the new location of data.

- Roles and responsibilities
  - PIs are responsible for maintaining data and metadata while data are in the PI's custody, and for implementing appropriate operational procedures to prevent an accidental loss of data
  - PIs are responsible for delivering the required set of Level 3 metadata to OE within 60 days following the conclusion of the OE expedition
  - The OE central repository and NOAA Data Centers are responsible for maintaining data after PIs have returned the data to NOAA. Through agreements, it is expected that discipline-specific research organizations will do the same
  - NOAA is responsible for maintaining a video library and management system that includes OE video data
  - NOAA and discipline-specific research organizations are responsible for implementing appropriate operational procedures to prevent an accidental loss of data
  - NOAA, in cooperation with discipline-specific research organizations, is responsible for implementing appropriate operational procedures to make data available to the research community and general public
  - OE will have oversight of the OE central catalog and central repository
  - OE is responsible for estimating the amount of data that will be collected each year
  - NOAA Data Centers and discipline-specific sites are responsible for providing appropriate infrastructure for maintaining OE data
  - PI, discipline-specific sites, and NOAA data repositories are responsible for notifying OE of any events that impact location or status of data sets that are in their custody
- Operational policies and procedures that need to be developed
  - Agreements that govern PI's provision of Level 3 metadata to OE within a specific time period and notification of OE when the status of data has changed

- Agreements with PIs regarding the time frame for returning data back to NOAA and protecting data against accidental loss
- Identification and implementation of applicable standards for metadata and data formats that may include specific metadata guidelines to facilitate standardization among the many data types and to help ease the management burden resulting from the large scope of data
- Back up and recovery procedures for data maintained under the OE program
- Policies and guidelines for data availability (e.g., what data will be available on-line and an operational data schedule)

#### 4.3.3.3 Data Access

There will be multiple access methods and delivery mechanisms for the data collected under the OE program depending on the category of users or data. Access to data includes the following steps:

- *Locating the data*. All data collected under the OE program will be described in the OE central catalog. A user can search the OE central catalog using multiple search criteria. The OE central catalog will specify the data's location and provide the user with information about data content, format, availability, and delivery mechanisms.
- *Selecting data sets*. Based on information obtained through a catalog search function, the user can select a subset of datasets returned by a search function.
- *Delivery of data*. Depending on the type, location, and availability of data, they may be available on-line for viewing and download, or by off-line delivery. NOAA may seek some cost recovery from users to defray the expense of data delivery.

Researchers working with discipline-specific research organizations can access data directly though their local infrastructure using access methods and delivery mechanism provided by the organizations with data custody, in addition to accessing data through the OE central catalog.

Dissemination of the OE data will be accomplished through the Internet using the OE or other data center Web sites. OE may develop multiple Web sites focused on various aspects of the OE program such as individual OE expeditions or specific scientific disciplines. Information disseminated through OE Web sites will be tailored to the various user communities (e.g., K-12 education, general public, or advanced users from the oceanographic research community).

- Roles and responsibilities
  - OE is responsible for overseeing the development of the OE central catalog

  - Individual data centers will be responsible for making data available to the user community
  - Host sites and discipline-specific research organizations will maintain and manage delivery mechanisms for OE data maintained by the distributed site
  - OE is responsible for developing OE Web sites with links to data centers and appropriate repositories
- Operational policies and procedures that need to be developed
  - Policies and procedures for disseminating OE information through OE Web sites
  - Policies for data access and availability for the OE data maintained by NOAA Data Centers and other OE data repositories

#### 4.3.3.4 Data Archive

NOAA is responsible for archiving all data collected under the OE program. OE will employ the capabilities of two NOAA Data Centers—NODC and NGDC—to achieve this goal. Depending on the data characteristic and user interest, data will be maintained in on-line, operational databases and after some period of time will be moved into the NOAA archive. Archived data will be available in different formats and media specified by the NOAA Data Centers. A nominal fee may be charged to defray the cost of delivery of archived data to the public.

- Roles and responsibilities
  - OE is responsible for categorizing data and determining a timeline for moving data from active databases to the NOAA archive
  - OE is responsible for managing the update of the OE central catalog as data are moved into the NOAA archive
  - OE is responsible for determining the data volume that needs to be archived every year
  - OE is responsible for determining formats and delivery mechanisms for archived data to meet the needs of the NOAA Data Centers
- Operational policies and procedures that need to be developed
  - Policies for moving OE data to the NOAA archive
  - Policies for delivery media to support data archival

## 4.4 Implementation Considerations

To facilitate the transition from this data management strategy to the implementation of a data management system, a set of actions must be undertaken to ensure an efficient and effective process. The development and implementation processes will require the establishment of partnerships within NOAA to support the level of desired capabilities discussed in Section 4.3, honor existing agreements and responsibilities for managing data,

and identify the requisite resources to design, build, and maintain the system. OE should consider the following list of actions in embarking on the data management implementation process. These actions encompass the specific recommendations contained in Section 4.3:

- Using the guidance contained in Sections 3 and 4 of this strategy, develop and publish an OE data management policy that can be reflected in Announcements of Opportunity and contract language and provide guidance to OE program participants on OE expectations concerning data management (Appendix F offers a strawman OE policy statement)
- Expand communications and seek a larger role for OE within NOAA in partnership with NOPP and other national-level government stakeholder groups to ensure consistency with standards and knowledge of emerging technologies and capabilities
- Designate a full time member of the OE government staff to be responsible for overseeing data management activities, policy development and enforcement, and guiding the development of the OE data management system
- Develop metadata guidelines that identify specific requirements for submission of metadata and include them in a guidance document that can be used by PIs and other program participants
- Communicate and coordinate anticipated archival requirements to NODC and NGDC to support new data management policies of OE and NOAA
- Establish a partnership with NESDIS to form an integrated project team to identify resources, develop an implementation plan, identify existing capabilities and assets that can be leveraged, and deliver a prototype OE central catalog and central repository for OE data as a step towards a distributed data management system
- Establish a partnership with NESIDS that includes the NOAA Central Library to form an integrated project team to identify resources, develop an implementation plan, and deliver a prototype NOAA Video Data Management System that satisfies OE data management needs
- Consider establishing an OE data management detail at a NESDIS data center to facilitate development of the catalog, repositories, and access to a video data management system during the initial peak effort period—approximately one year—prior to transitioning to a routine level of effort

# APPENDIX A VIDEO DATA MANAGEMENT SOLUTIONS

A component of the information gathering completed in support of this data management strategy was an investigation of current and emerging video data management technologies. Site visits were made to several locations involved in the management of video data with oceanographic themes to ascertain existing practices and technologies and identify those with potential applicability to a centralized NOAA video data management system.

An extensive digital video data management program is in place to support exploration and scientific research at the Monterey Bay Aquarium Research Institute (MBARI). MBARI invests in state-of-the-art equipment and has developed its own tailored video data management software. Most of the video data gathered from ROV platforms deployed from their vessels is maintained internally for a two-year period to allow MBARI staff exclusive access. Distribution of video data to outside sources is accomplished primarily via requests received through the public MBARI web site, and is distributed in multiple formats including optical disks or on-line compressed files. Collection, storage, and archival of video data is performed using annotated videotapes. Whereas video from cameras were once recorded on analog tapes, now all video is recorded and archived on Digital BetaCam, Mini-DV, and most recently onto HDTV medium to support the companion camera on one of the MBARI ROVs. All of the videotapes are stored and archived in a dedicated, temperature and humidity controlled room with moving storage shelves. MBARI is beginning to investigate archival of data on DVD to increase storage life and reduce space requirements. Video annotations are constructed by a dedicated video editor in the video laboratory using the original media, supporting hardware and software, and the audio record made by participating scientists and data managers during data collection. The MBARI video laboratory includes seven Digital Betacam decks, three mini-DV decks, and two HDTV decks. It also hosts a video information management system (VIMS) and is linked to an internal MBARI relational database so that annotation files, video metadata, designated video clips, and frame captures are universally accessible by the MBARI scientific staff via a Web-based front end. As a component of VIMS, MBARI has developed an in-house, developmental software package known as the Video Information Capture and Knowledge

Inferencing (VICKI) system. The video editor in the laboratory employs an intuitive, icon-based interface in VICKI to digitally annotate video data, identify and tag unique biota, geological features, and other significant objects in the video stream, and capture relevant frame grabs and video clips for storage on the central relational database. Navigation through the video data and annotation files is performed using time increment searches. The VICKI capability has significant potential for expansion and application by other organizations involved in oceanographic exploration and research. Currently video data are annotated during post-analysis in the video laboratory; however, a MBARI goal is to provide on-scene scientists and data managers with a VICKI capability to support real-time annotation while viewing data on ROV control room monitors as it is being collected. A microwave link is used to establish a live video connection between ROVs operating from deployed vessels in Monterey Bay and both MBARI and a public theater at the Monterey Bay Aquarium, with the potential to support real-time annotation from the MBARI video laboratory. VIMS developers are also acquiring database technology that will assist in populating the external MBARI Web page. All unique MBARI data, including video data, are collected on servers unique to particular data types and are accessible by the central relational database. This architecture is driven by a MBARI focus on knowledge management, with a concurrent focus on improving data cataloging and metadata employment. The MBARI relational database employs a software program called HARVEST, which is used by in-house staff to associate stored data from previous missions with currents tasks to identify common elements and trends.

Video data management practices were examined at the Discovery Channel headquarters in Bethesda, Maryland. Discovery Channel personnel at their headquarters operations described their organization as a worldwide video media company that captures and manages imagery rather than creating it. They routinely manage the storage, access and retrieval of video data. The television broadcast component of the company performs quality control operations on the video data they process; however, the quality control is performed on all metadata at the headquarters site. Each contractor they work with conforms to a "style guide" that specifies all data requirements and operations to be followed. Production efforts are performed at locations around the world. Research efforts are directed towards still and moving footage

and acquiring and managing the content and format of information. Software tools from Convera Corporation (www.convera.com) and Virage Corporation (www.virage.com) are employed for video logging and indexing. Additional ongoing work is focused on the application of metadata fields for photos for access via the Internet. There is also work related to employing a natural language search capacity, where searches of sentences, verbs, and nouns are performed. Search engines in use include the freeware software packages Freenet™ and Wordnet™ (used within the photo library). They have adopted a video-cataloging manual based on the machine readable cataloging (MARC) format. They use this format for bibliographic information that is arranged in a prescribed format and on a prescribed medium—such as magnetic tape—that allows information to be read by electronic data processing equipment. The Discovery Channel is experimenting with a variety of new video indexing techniques that include the conversion of videotape data collected in the field to video compact disk, super video CD (SVCD), and DVD formats. They are also using new time coding techniques to map events to time-referenced images. While Society of Motion Picture and Television Engineers (SMPTE) formats of 30 frames per second are employed for most work, they are examining very low frame rates (down to 2 frames/second) for deep ocean sea life recognition work. Most digital video in use is relatively low resolution data (300-500 kilobytes per second). A goal is to combine metadata with video at the time of creation. They are also examining digital video compression techniques in all high storage and data transfer operations. Other tools used for video processing and metadata classification include the following:

- Artesia (http://www.artesia.com), for digital asset management
- eMotion (http://www.emotion.com) for digital media management
- Media 360 Assential, from Assential Software Corporation, for database applications
- Informix Media360 (http://www.ibm.com) for information management solutions
- Northplains Software Corp. (http://www.northplains.com) eVision tool for visual search
- Techmath Corp. (Germany) for digital visualization tools
- Bulldog (http://www.buldog.com), a digital asset management tool

The National Geographic headquarters in Washington, DC has a long history of using digital video, from its inception. This organization has had a unique partnership with NOAA over

the past five years conducting research in and promoting the twelve U.S. National Marine Sanctuaries through the Sustainable Seas Expeditions (SSE). A significant part of the SSE was the employment of new and innovative video technologies in the ocean environment. For collection, the SSE employed Sony Mini-GVD300 recorders in their manned submersibles along with an audio channel dedicated to a microphone used by the submersible operator. The GVD300 provides 550 horizontal lines of resolution. The SSE video suite was well equipped not only to record images on site, but also to edit and manipulate images while at sea. Four digital video decks were used to duplicate and transfer video data, one Beta SP recorder was employed for archiving, a Macintosh G3 system provided desktop editing services, and an Epson Photo EX printer produced hard copy images on-scene. Desktop editing was accomplished using Edit-DV software and Adobe PhotoShop. Video image frame captures could be processed, edited, and electronically transferred in less than an hour. As with the Discovery Channel, National Geographic uses the Convera software package for data asset management. Convera's Screening Room™ application is used to capture and display video, while their Retrieval Ware™ package is used for concept and text searching. Recording equipment employed by National Geographic are mostly of the Beta SP and Omni DV types. One unique tool known as CritterCam is used to tag and track biota such as marine mammals and incorporates continuous metadata collection on its recording media, including temperature, salinity, and depth information. Digital video from all over the world is sent to the National Geographic's headquarters in Washington, DC. These data are processed, assigned a unique production reference identifier, and then key frames and annotated time codes are extracted. Metadata fields are populated into databases such as Oracle™ using SQL. These processed images, with metadata, are forwarded to a central image storage facility in Portland, Oregon, where they can be retrieved via specialized software tools designed around a Web browser. This browser is essentially an internal Internet search engine called the Digital Archive. Image files are transmitted using a virtual private network (VPN), 100 Base-T Fast Ethernet lines, or over the Internet for smaller files. The central storage facility in Portland currently maintains two servers of four terabytes capacity each. One is used for storing images and the other for storing database and textual information. The Digital Archive employs Microsoft Windows™ Media (with ASF files) for streaming video. A decision was made to bypass MPEG-1, which is the first generation of a widely used video

compression standard. Instead, the systems employs MPEG-2 with plans to migrate to the next generation known as MPEG-4. MPEG-2 is a video compression standard that compares each successive frame of video and records only the changes from the previous frame. This greatly reduces bandwidth requirements and allows MPEG-2 to carry more information per unit of time. The resultant free bandwidth can be exploited for improved image quality or additional data, a requirement for some fast action video applications on HDTV format. Most DVD players and satellite television broadcasting services currently use MPEG-2. National Geographic has historically had limited involvement with quality assurance techniques but is increasing its attention and investment in this area.

# APPENDIX B MIS DEVELOPMENTS IN NOAA

There is a wide range of potential commercial solutions that could be adapted with a moderate commitment of resources to meet emerging OE MIS requirements. Two MIS solutions are under development within NOAA that are tailored to the management of federally sponsored oceanographic research information. These programs were examined during the course of the development of this data management strategy.

The NOS Special Projects Office is sponsoring an applicable MIS development. A prototype was tested during the Islands in the Stream Expedition and the Sustainable Seas Expedition in 2001. It employs a FileMaker Pro® database system on a desktop personal computer and is designed for direct use during the conduct of at-sea operations. Its principal focus is the management of Level 1 metadata during and after an expedition. Strengths include the extensive template of daily operations, activities, and dive information, situation report generation, and ability to represent a comprehensive summary of the detailed activities of an expedition. Its primary weakness is its reliance on a dedicated, on-scene data manager to continuously interact with expedition participants and capture the requisite information. While there are considerable advantages to dedicating manpower for these purposes, such resources may not be consistently available. This system does not yet include an ability to manage contacts and programmatic-level information; however, the current developmental version was not designed to provide these functions.

The second MIS development is being sponsored by the National Undersea Research Program (NURP) and is primarily designed as a means for the regional NURP centers to measure and report on the progress of ongoing research activities. It is a distributed system that allows multiple sites access via the Internet to a central relational database that tracks a wide variety of entities related to sponsored research projects, including investigator administrative information, cost and performance data, Level 1 metadata related to research expeditions, and representative samples of data including digital imagery and short video clips. Its key strength is its ability to provide managers with information necessary to provide oversight of research awards throughout the solicitation, peer review, execution, and

reporting process. It is not currently designed to capture detailed Level 1 metadata during ongoing afloat operations like the NOS system, but could be expanded to include this capability. Its implementation would require some level of technical support to maintain the relational database and Internet connectivity.

# APPENDIX C MARINE GEOLOGY AND GEOPHYSICS WORKSHOP RECOMMENDATIONS

On May 14-16, 2001 the National Science Foundation and the Office of Naval Research sponsored a workshop on *Data Management for Marine Geology and Geophysics: Tools for Archiving, Analysis, and Visualization*.[43] The workshop's objective was to bring together researchers, data collectors, data users, engineers, and computer scientists to assess the state of existing data management efforts in the marine geology and geophysics (MG&G) community, share experiences in developing data management projects, and help determine the direction of future efforts in data management. The workshop agenda was organized around presentations, plenary discussions, and working group discussions. The presentations provided examples of the needs of data users, the needs of large, multidisciplinary MG&G projects, existing data management projects in the community, tools that have been developed for data access and analysis, examples of organizations with centralized databases, and current topics in information technology. Working groups addressed questions concerning three different themes: (1) the structure of a data management system, (2) data archiving and access, and (3) data documentation. The working groups were also asked to recommend strategies to permit MG&G data management to move forward in these areas.

On the structure of a data management system:

- Create permanent, active archives for all MG&G data
- Manage data using a distributed system with a central coordinator
- Manage different data types with user-defined centers
- Support area or problem specific databases if scientifically justified, but these databases should link to rather than duplicate data holdings within discipline specific data centers
- Evaluate the data management system using oversight and advisory committees, in-depth peer reviews at renewal intervals, and *ad hoc* panels to access each data center's contribution to science
- Fund core operating costs of the distributed data centers

On the data archiving and access process:

- Always archive raw data. Archive derived data for high demand products
- Store data in open formats

- Develop standardized tools and procedures to ensure quality at all steps from acquisition through archiving
- Improve access to common tools for data analysis and interpretation for the benefit of the community
- Build data centers to address the needs of a diverse user community, which will be primarily scientists
- Enforce timely data distribution through funding agency actions
- Promote interactions among federal agencies and organizations, and international agencies, to define data and metadata exchange standards and policies

On data documentation:

- Create a centralized and searchable on-line metadata catalog
- Require ship operators and principal investigators to submit Level 1 metadata and cruise navigation to the centralized metadata catalog at the completion of each cruise as part of the cruise reporting process
- Generate a standard digital cruise report form and make it available to all chief scientists for cruise reporting (Level 2 metadata)
- Require individual principal investigators to complete and submit standard forms for Level 1 and 2 metadata for field programs carried out aboard vessels not in the UNOLS fleet
- Generate a standardized suite of Level 1 and 2 metadata during operation of seafloor observatories and other national facilities, and submit to the central metadata catalog
- Require Level 3 metadata within each discipline specific data center. Archiving of publications related to the data should also be included (Level 4 metadata)
- Follow nationally accepted metadata standards (particularly for Levels 1 and 3 metadata)

The workshop participants identified that a clear top priority was to define and establish a centralized metadata catalog. The metadata catalog should be broad, containing information on as many data types as possible. It should support geospatial, temporal, keyword, and expert level searches of each data type. The catalog should be a circular system that allows feedback from the user to the originator. The metadata catalog should serve as the central link to the distributed network of data centers where the actual data reside.

The workshop further concluded that the construction of a central metadata catalog for Levels 1 and 2 metadata was viewed with the highest priority. Level 1 metadata should be generated during data acquisition and should be submitted to the central metadata archive

immediately following a field program. Level 2 metadata should also be archived within the central metadata catalog, whereas Level 3 metadata would reside with the actual data themselves. The requirements for Levels 1 and 2 metadata can be standardized whereas Level 3 metadata requirements will vary by data type.

# APPENDIX D OCEAN EXPLORATION DATA FORMATS

Digital data for oceanographic exploration and research are collected using a variety of instruments and sensors at different sampling rates and resolutions. A general primer on fundamental oceanographic data formats is available from the National Centers for Atmospheric Research (NCAR).[42] These data are typically stored and transported on a variety of media selected by the cognizant PI. Depending on the collection equipment employed, media may include handwritten forms, magnetic media such as tapes and removable or internal computer disks, or optical media such as writeable compact disks (CDs) and digital versatile discs (DVDs).

Common scientific formats used for these oceanographic data that are widely employed by oceanographers include textual formats such as the American Standard Code for Information Exchange (ASCII), ASCII Common Separated Variable (CSV), Hierarchical Data Format (HDF), Common Data Format (CDF), and Network Common Data Format (netCDF). Imagery data are typically recorded in raster file formats that include the Tagged Image File Format (TIFF), Graphic Interchange Format (GIF), JPEG format, and bitmap (BMP). Vector graphics file formats include PostScript, encapsulated PostScript (EPS), and Drawing eXchange Format (DXF). The most commonly employed graphics metafile formats include Computer Graphics Metafile (CGM), Windows™ Metafile Format (WMF), Graphical Environment Manager (GEM), and Desktop Color Separation (DCS). Digital tape formats typically consist of standards recognized by the Society of Exploration Geophysicists (SEG) Technical Standards Committee.[43] Acoustic files may be in a binary format with conversions to formats common to Internet users such as Waveform Audio (WAV) and MPEG Layer 3 (MP3). Video data media include Video Home System (VHS), Super VHS (SVHS), Hi8™ and Video8™, Digital Video (DV) and miniDV, Digital Video Camera (DVCAM), Video CD, DVD, and HDTV. Also, the recent surge of multimedia applications for desktop computers has led to the development of a variety of video compression/decompression (CODEC) methods, many of them proprietary, for use by desktop digital video implementations. Among the most commonly CODEC methods currently employed for oceanographic data are MPEG, Digital Visual Interface (DVI), and Cinepak™. Common

gridded data formats include Geophysical Data Base (GDB) and Gridded Binary (GRIB), although there are a larger number of tailored formats for bathymetry and hydrography that have been adopted by users of these data, such as NGDC's Earth Topography – 2 minute (ETOP02) and the Navy's Digital Bathymetric Database (DBDB). The International Hydrographic Organization (IHO) alone sanctions a set of data format standards that include S-44, S-57, Feature and Attribute Coding Catalogue (FACC), Regional Engineering and Environmental GIS (REEGIS) format, Digital Geographic Information Exchange Standard (DIGEST), Tri-Service Spatial Data Standard (TSSDS), and an emerging National Hydrographic Data Content Standard for Coastal and Inland Waterways format that is compatible with prior standards and is being designed under the auspices of the FGDC to be compatible with most GIS software.

# APPENDIX E A GIS-BASED TAXONOMY TEMPLATE FOR OCEAN EXPLORATION DATA

The information contained in Table E-1 provides a template for the development of a data taxonomy for ocean exploration data that supports the use of GIS tools for analysis, information discovery, and display. The list of data types is not inclusive but is provided to illustrate a cross section of types and likely attributes of these types that could be exploited by a GIS. The attribute names are intuitive. The topology represents the spatial relationship between connecting or adjacent features in a GIS coverage. The list of data types and associated attributes is expected to expand and be refined as data modeling efforts associated with the engineering of a companion data management system are undertaken. Additional information supporting this type of scientific information modeling may be found in manuscripts describing modeling efforts made in conjunction with the NOAA VENTS program.[44]

**Table E-1. Data Taxonomy Template**

| **DATA TYPE:** Sidescan/Multibeam Sonar | | **TOPOLOGY:** Line, Polygon | | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | instrument_ID | sample_rate | channels |
| | origin_LAT | origin_LON | end_LAT | end_LON |
| | location | date | origin_time | end_time |
| | beam_angle | ping_number | frequency | band_width |
| | pulse_width | range_scale | ship_telemetry | fish_telemetry |
| | fish_depth | fish_elevation | | |

| **DATA TYPE:** Nav/Dive/Trackline Log | | **TOPOLOGY:** Point, Line | | |
|---|---|---|---|---|
| **ATTRIBUTES:** | Vehicle/tow_name | Location | Sublocation | origin_LAT |
| | origin_LON | end_LAT | end_LON | origin_depth |
| | min_depth | max_depth | end_depth | date |
| | origin_time | end_time | pilot_ID | observer_ID |
| | surface_obs | geomorph_ID | geo_sediment_ID | habitat_ID |
| | biota_ID | human_impact_ID | samples_ID | transect_ID |

| **DATA TYPE:** Hydrophone | | **TOPOLOGY:** Point | | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | instrument_ID | sample_rate | LAT |
| | LON | depth | frequency | orientation |
| | channels | array_configuration | geoevent_ID | species_ID |
| | event_time | magnitude | duration | |

| **DATA TYPE:** Video | | | **TOPOLOGY:** Point, Line | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | camera_type | media_format | date |
| | origin_LAT | origin_LON | origin_depth | origin_time |
| | end_LAT | end_LON | end_depth | end_time |
| | EM_band | content_ID | | |

| **DATA TYPE:** Still Imagery | | | **TOPOLOGY:** Point | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | camera_type | media_format | date |
| | time | LAT | LON | depth |
| | EM_band | content_ID | | |

| **DATA TYPE:** Core | | | **TOPOLOGY:** Point | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | core_type | date | time |
| | LAT | LON | depth | penetration |
| | sediment_ID | biota_ID | sediment_chemistry | |

| **DATA TYPE:** CTD | | | **TOPOLOGY:** Point, Line | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | instrument_ID | cast_number | date |
| | time | LAT | LON | max_depth |
| | sample_rate | conductivity | temperature | pressure |
| | salinity | water_sample_ID | water_chemistry | particulates |
| | transmissivity | | | |

| **DATA TYPE:** ADCP | | | **TOPOLOGY:** Line, Polygon | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | instrument_ID | date | average_depth |
| | LAT | LON | azimuth | telemetry |
| | bottom_track | origin_time | end_time | frequency |
| | pulse | current_velocity | current_dir (u) | current_dir (v) |
| | current_dir (w) | | | |

| **DATA TYPE:** Echosounder | | | **TOPOLOGY:** Line, Polygon | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | instrument_ID | date | average_depth |
| | origin_LAT | origin_LON | origin_depth | origin_time |
| | end_LAT | end_LON | end_depth | end_time |
| | frequency | pulse | sounding_depth | biota_distribution |
| | biomass | | | |

| **DATA TYPE:** Biological/Geological Samples | | | **TOPOLOGY:** Point | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | instrument_ID | sample_ID | date |
| | time | LAT | LON | depth |
| | composition | physiology | preservation_tech | disposition |

| **DATA TYPE:** Trawl | | | **TOPOLOGY:** Line | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | trawl_ID | date | average_depth |
| | origin_LAT | origin_LON | origin_depth | origin_time |
| | end_LAT | end_LON | end_depth | end_time |
| | biota_distribution | biomass | samples_ID | |

| **DATA TYPE:** Seismometer | | | **TOPOLOGY:** Point | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | station_ID | instrument_ID | LAT |
| | LON | depth | event_date | event_start_time |
| | event_end_time | sample_rate | event_phase | event_magnitude |

| **DATA TYPE:** Traps | | | **TOPOLOGY:** Point | |
|---|---|---|---|---|
| **ATTRIBUTES:** | filename | trap_ID | deploy_date | deploy_time |
| | recover_date | recover_time | LAT | LON |
| | depth | samples_ID | | |

# APPENDIX F STRAWMAN OE DATA MANAGEMENT POLICY

The following strawman OE data management policy was adapted from the policy guidance provided by the NSF Ridge Interdisciplinary Global Experiments (RIDGE) Endeavour Segment Seafloor Observatory Project, which is based on US Joint Global Ocean Flux Study (JGOFS) data policy. It integrates the policy recommendations contained within this data management strategy and represents a recommended data management policy statement for OE.

## INTRODUCTION

This data management policy is designed to address the needs of both the NOAA ocean exploration program and individual investigators. Central to this policy is timely submission and sharing of all data collected during exploration activities under the auspices of the NOAA Office of Ocean Exploration (OE). A strong commitment to data management is required of each participating Principal Investigator (PI). In accepting full or partial sponsorship by the government via OE, each PI agrees with the obligation to meet the following suite of data management requirements as an integral aspect of their participation in the program. All proposals to participate in the OE program must include resource needs and a description of the associated application of those resources necessary to comply with all aspects of this data management policy. The level of PI compliance with this policy will be monitored, recorded, and included as a measure of performance during reviews of any new proposals for work in subsequent field work cycles.

NOAA policy on the release of marine environmental data to the public domain is clear with regard to its submission to NODC and NGDC, as appropriate, for archival. Thus, PIs participating in OE program activities share these obligations. To paraphrase NOAA policy, recipients of federal funding supporting collection of marine environmental data must release these data to the NODC (for example, ocean physical data, ocean chemical data, ocean biology data) or the NGDC (for example, geophysical, geological and geochemical data) within one year of the date of collection. Additionally, NOAA guidelines state that post-cruise inventory information be completed at the end of a cruise or other exploration activity.

This inventory information is provided in two ways: in the form of a Cruise Report generated by the Chief Scientist and in the form of data set object and access information provided by the PI as metadata that includes data formats, quality assurance procedures, other processing such as the application of calibration or compression techniques, and any other elements necessary to describe the content, format, and accessibility of the collected data set. All initial submissions of, and subsequent modifications to, metadata shall be made in Federal Geographic Data Committee (FGDC) compliant format. OE will specify the minimum set of metadata attributes required from PI's to support the initial submission.

To facilitate data management, a data management system (DMS) will be implemented, maintained, and operated by an OE Data Manager (OEDM). The role of the OEDM will be to ensure that all OE program data sets are readily accessible and that tools are available to access all data contained in the system on a common time base and within a common spatial framework.

Ultimately, all data will be archived at the NODC or the NGDC in accordance with NOAA policy.

OCEAN EXPLORATION PROGRAM DATA MANAGEMENT POLICY

This OE program data management policy is predicated on openness and sharing of exploration data for the mutual benefit of all exploration stakeholders, balanced by recognition of the PI's individual rights to these data. This policy sets responsibilities for release of data with the understanding that some collected data will require analytical or data reduction procedures that prevent immediate release after collection of samples or retrieval of instruments.

PIs will provide the OEDM with access to all raw OE program data and data sets derived from these raw data as soon as practical following collection, but in any case not later than one year following the date of collection. OE will consider limited restrictions at the request of the PI on the use of data during this one-year time period. PIs will provide the OEDM access to these data sets by hosting them on Internet-accessible data servers at host site

locations or discipline-specific research organizations that satisfy minimally acceptable standards established by the OE for data security, throughput, and other considerations. As an alternative, PIs may satisfy this access requirement by forwarding copies of data sets and accompanying metadata to the OEDM. The OE understands that, in rare cases, data sets will require lengthy analytical or processing procedures. When situations exist that prevent the PI from providing timely access to data, the status of the data should be acknowledged by the PI through the submission of updated metadata to the OEDM so that data stakeholders are aware of the data's location, condition, and accessibility.

While recognizing the legitimate rights of data originators to the first use of the data they collect, the OE policy is that availability to ocean exploration data should be restricted only in exceptional cases. Data normally becomes publicly available for use without restriction one year after origination, and sooner if facilitated by agreement between OE and the PI. All data users will be expected to properly acknowledge the source and sponsor of the data, whether or not restrictions apply to its use.

The following series of responsibilities for PIs, Chief Scientists, and the OEDM result from the above principles.

RESPONSIBILITIES OF THE CHIEF SCIENTIST

1. The Chief Scientist of each OE sponsored exploration activity shall maintain a detailed operations log for every sampling operation during a given cruise or leg of a field collection effort. This log shall include all information necessary for creating a Cruise Report and will include the following minimum elements as appropriate: date, location, vessel identification, operating status, participating personnel, operations and safety information, submersible, ROV, and AUV dive number, operation number, and track lines, station number, transect description, dive objectives, instruments employed, and data inventory, ancillary operations information including conductivity-temperature-density (CTD) casts and other instrument deployments, education and outreach information, summaries of significant accomplishments and new discoveries, and other comments as appropriate

2. The Chief Scientist will direct participating PI's and any personnel with on-scene data management responsibilities to provide Level 1 metadata necessary to support subsequent generation of the Cruise Report.
3. The Chief Scientist will submit the Cruise Report to OE and the OEDM within 90 days of the end of the exploration activity.

RESPONSIBILITIES OF THE PRINCIPAL INVESTIGATOR

1. In response to OE Announcements of Opportunity, PIs will identify within their proposals the required resources and procedures that will be employed to satisfy the requirements contained in this OE program data management policy.
2. PIs will satisfy the data management provisions contained in contract awards from OE in response to their proposals, in accordance with the provisions of this policy.
3. During exploration activities, PIs will provide the Chief Scientist and supporting on-scene data managers with Level 1 metadata necessary to support post-expedition cruise reporting.
4. PIs will submit FGDC-compliant Level 3 metadata for the data sets under their cognizance to the OEDM not later than 90 days following the completion of the data collection activity. OE will provide guidance containing the minimally acceptable list of metadata elements to support this initial submission. PIs are encouraged but not required to use automated metadata generation tools to develop compliant metadata. Level 3 metadata shall also include the following:
   a) Quality assurance and calibration procedures, with statements that convey the limitations of associated data based on these procedures
   b) Guidance to potential users that reflect their responsibility for the data's use or misuse in further analyses or comparisons, that the federal government does not assume any liability to users or third parties, and that the government will not indemnify users for liability due to any losses resulting from the use of the data
   c) Citation instructions for potential users
5. PIs will submit changes and additions to Level 3 metadata as necessary to reflect the existence of additional derived data sets, changes in storage location, modified access paths, and any other changes that should be reflected in the OE central catalog.

6. PIs will secure collected data against possible loss through appropriate backup and recovery procedures contained in OE guidance.
7. PIs will provide the OEDM access to the data sets under their cognizance as soon as is practical and in any case not later than one year following the date of collection, unless an extension is specifically granted by OE. PIs may provide this access to data using either of the following methods (in priority order of OE preference):
    a) Hosting data on Internet-accessible servers at host site locations or discipline-specific research organizations that satisfy minimally acceptable standards established by the OE for data security, throughput, and other considerations; data must be reachable via search capabilities at the OE central catalog
    b) Forwarding copies of data sets and accompanying metadata to the OEDM for hosting on the OE central repository
8. PIs will ensure host sites and discipline-specific research organizations maintain and manage the delivery infrastructure for their OE sponsored data held at distributed sites.
9. PIs will identify any desired data access restrictions to the OEDM and rationale for the desired restrictions.
10. PIs are responsible for the quality and correctness of the data available for access via the DMS, and will intervene to correct data quality issues upon notification.

RESPONSIBILITIES OF THE OE DATA MANAGER

1. The OEDM will regulate and enforce OE data management policy guidance, provide oversight of the data management process from collection planning through archival, and communicate regularly with PIs and other OE program participants.
2. The OEDM will manage the operation of a secure, Web-based OE central catalog and portal, and will ensure all metadata for data collected during exploration activities are included in the catalog and made accessible and searchable by data users
3. The OEDM will manage the operation of a secure OE central repository for government-owned data and other data provided by PIs that will not be hosted at distributed sites.
4. The OEDM will coordinate with NOAA/NESDIS to manage, maintain, and improve the capabilities of the OE central catalog and repository.
5. The OEDM will coordinate with NOAA/NESDIS to ensure storage, access, and archival of imagery and video data from exploration activities.

6. The OEDM will provide NOAA/NESDIS with access to OE program metadata to support access to the OE central catalog via other oceanographic data catalogs and clearinghouses.
7. The OEDM will forward copies of raw and derived data sets as they become accessible to NODC or NGDC, as appropriate, for archival.
8. The OEDM will oversee the conduct of quality assessment of all data and will notify PIs of any problems identified in their data sets.
9. On behalf of OE and when required, the OEDM will designate an on-scene exploration data manager to participate in selected exploration activities and assist the Chief Scientist and PIs in complying with metadata requirements.

## LIST OF ACRONYMS

| | |
|---|---|
| **AMIA** | Association of Moving Image Archivists |
| **ANSI** | American National Standards Institute |
| **ASCII** | American Standard Code for Information Interchange |
| **AUV** | autonomous underwater vehicle |
| **BMP** | bitmap |
| **CD** | compact disk |
| **CDF** | common data format |
| **CGM** | computer graphics metafile |
| **CLASS** | Comprehensive Large Array-data Stewardship System |
| **CO** | Commanding Officer |
| **COTS** | commercial off-the-shelf |
| **CODATA** | Committee on Data for Science and Technology |
| **CODEC** | compression/decompression |
| **CONOPS** | concept of operations |
| **CORE** | Consortium for Oceanographic Research and Education |
| **CSDGM** | Content Standard for Digital Geospatial Metadata |
| **CSR** | Cruise Summary Report |
| **CSV** | comma separated variable |
| **CTD** | conductivity-temperature-density instrument |
| **DBDB** | Digital Bathymetric Database |
| **DBMS** | Database Management System |
| **DC** | Discovery Channel |
| **DCS** | desktop color separation |
| **DEI** | data exchange interface |
| **DIGEST** | Digital Geographic Information Exchange Standard |
| **DODS** | Distributed Oceanographic Data System |
| **DTD** | Document Type Declaration |
| **DV** | digital video |
| **DVCAM** | digital video camera |
| **DVD** | digital video disc or digital versatile disc |
| **DVI** | digital visual interface |
| **DXF** | drawing exchange format |
| **EDL** | edit decision list |
| **EPS** | encapsulated postscript |
| **ETOPO2** | earth topography – 2 minute |
| **FACC** | Feature and Attribute Coding Catalogue |
| **FGDC** | Federal Geographic Data Committee |
| **FIPS** | Federal Information Processing Standards |

| | |
|---|---|
| **FOIA** | Freedom of Information Act |
| **FY** | fiscal year |
| | |
| **GDB** | Geophysical Database |
| **GEM** | Graphical Environment Manager |
| **GIF** | graphic interchange format |
| **GIS** | geographic information system |
| **GRIB** | gridded binary |
| | |
| **HDF** | hierarchical data format |
| **HDSA** | High Density Storage Association |
| **HDTV** | high definition television |
| **HTML** | hypertext markup language |
| | |
| **ICES** | International Council for the Exploration of the Sea |
| **ICSU** | International Council for Science |
| **IEC** | International Electrotechnical Commission |
| **IEEE** | Institute of Electrical and Electronics Engineers |
| **IHO** | International Hydrographic Organization |
| **IOC** | Intergovernmental Oceanographic Commission |
| **IPR** | intellectual property rights |
| **ISO** | International Organization for Standardization |
| | |
| **JGOFS** | Joint Global Ocean Flux Study |
| **JPEG** | Joint Photographic Experts Group |
| **JSL** | Johnson-Sea-Link |
| | |
| **MARC** | machine readable cataloging |
| **MBARI** | Monterey Bay Aquarium Research Institute |
| **MG&G** | Marine Geology and Geophysics |
| **MGD77** | Marine Geophysical Data Exchange Format 77 |
| **MIS** | management information system |
| **MOOS** | MBARI Ocean Observing System |
| **MP3** | MPEG Layer 3 |
| **MPEG** | Moving Picture Expert Group |
| | |
| **NARA** | National Archives and Records Administration |
| **NBII** | National Biological Information Infrastructure |
| **NCDDC** | National Coastal Data Development Center |
| **NCITS** | International Committee for Information Technology Standards |
| **NESDIS** | National Environmental Satellite, Data, and Information Service |
| **netCDF** | network common data format |
| **NGDC** | National Geophysical Data Center |
| **NIST** | National Institute of Standards and Technology |
| **NMAS** | National Map Accuracy Standards |
| **NMFS** | National Marine Fisheries Service |

| | |
|---|---|
| **NOAA** | National Office of Atmospheric Administration |
| **NODC** | National Oceanographic Data Center |
| **NOPP** | National Oceanographic Partnership Program |
| **NOS** | National Ocean Service |
| **NRC** | National Research Council |
| **NSDI** | National Spatial Data Infrastructure |
| **NSF** | National Science Foundation |
| **NURP** | National Undersea Research Program |
| | |
| **ODMG** | Object Data Management Group |
| **OE** | Office of Ocean Exploration |
| **OEDM** | Ocean Exploration Data Manager |
| **OMB** | Office of Management and Budget |
| **ONR** | Office of Naval Research |
| | |
| **PI** | principal investigator |
| **PMEL** | Pacific Marine Environmental Laboratory |
| | |
| **RDA** | remote database access |
| **REEGIS** | Regional Engineering and Environmental GIS |
| **RF** | radio frequency |
| **RFC** | request for comment |
| **RIDGE** | Ridge Interdisciplinary Global Experiments |
| **ROSCOP** | Report of Observations/Samples collected by Oceanographic Programmes |
| **ROV** | remotely operated vehicle |
| **RV** | research vessel |
| | |
| **SAIF** | spatial archive interchange format |
| **SCS** | science computer system |
| **SDTS** | Spatial Data Transfer Standard |
| **SEG** | Society of Exploration Geophysicists |
| **SMPTE** | Society of Motion Picture and Television Engineers |
| **SQL** | structured query language |
| **SSE** | Sustainable Seas Expedition |
| **SVCD** | super video CD |
| **SVHS** | super video home system |
| | |
| **TB** | terabytes |
| **TCP/IP** | transmission control protocol/Internet protocol |
| **TIFF** | tagged image file format |
| **TSSDS** | Tri-Service Spatial Data Standards |
| | |
| **UNESCO** | United Nations Educational, Scientific, and Cultural Organization |
| **UNOLS** | University-National Oceanographic Laboratory System |
| **USGS** | U.S. Geological Survey |
| **UTC** | universal time coordinates |

| | |
|---|---|
| **VCD** | video compact disk |
| **VDMS** | video data management system |
| **VHS** | video home system |
| **VICKI** | Video Information Capture and Knowledge Inferencing |
| **VIMS** | video information management system |
| **VPN** | virtual private network |
| | |
| **W3C** | Worldwide Web Consortium |
| **WAV** | waveform audio |
| **WIPO** | World Intellectual Property Organization |
| **WMF** | windows metafile format |
| | |
| **XML** | extensible markup language |

## ENDNOTES

[1] *Discovering Earth's Final Frontier: A U.S. Strategy for Ocean Exploration*. Report of the President's Panel on Ocean Exploration, University Corporation for Atmospheric Research, October 10, 2000.

[2] *Ibid.,* page 6.

[3] Briscoe, Melbourne G., *Ocean Exploration in the U.S. Navy*. Presentation to the President's Panel on Ocean Exploration, Washington, DC, August 20, 2000, page 4.

[4] Extracted from the NOAA 2002 budget submission developed by Mr. Andrew Larkin, NOAA Office of Legislative Affairs, 202.482.4630.

[5] McNutt, Marcia, K., *Ocean Exploration.* Presentation at the Third Annual Roger Revelle Commemorative Lecture, National Academy of Sciences Auditorium, Washington, DC, November 1, 2001, page 16.

[6] *Op. Cit., Discovering Earth's Final Frontier: A U.S. Strategy for Ocean Exploration*.

[7] *Ocean Exploration Program Strategic Framework*. Office of Ocean Exploration, National Oceanic and Atmospheric Administration, published by Mitretek Systems, Inc., June 24, 2002.

[8] *Ibid.,* page 9.

[9] *A Digital Strategy for the Library of Congress.* Report of the Committee on an Information Technology Strategy for the Library of Congress, National Research Council, National Academy Press, LC 00-111489, 2000.

[10] Wright, D.J. et al., *A Scientific Information Model for Deepsea Mapping and Sampling*. Marine Geodesy, 20(4): 1997, pp.367-379.

[12] Stinus, et al., *The New National Coastal Data Development Center: A Status Report.* American Meteorological Society Annual Meeting, Orlando, FL, January 13-17, 2002.

[14] The Internet Request for Comments (RFC)-1006 is a guide used for executing RDA over a TCP/IP connection.

[15] The OpenGIS™ Abstract Specification, The OpenGIS™ Consortium, Inc., Document Number 99-100r1.doc, 1999; http://www.opengis.org/

[16] *Generation of Ocean Exploration Propelled by High-Speed Wireless Technology*. Office of Legislative and Public Affairs press release, National Science Foundation, December 10, 2001, PR 01-102.

[17] Michener, W. et al, *Nongeospatial Metadata for the Ecological Sciences*. Ecological Applications, 7 (1), 1997, pp. 330-342.

[18] *Data Management for Marine Geology and Geophysics*. Workshop Report, La Jolla, CA, May 14-16, 2001; http://www.geo-prose.com/projects/pdfs/data_mgt_report.low.pdf.

[19] http://www.ices.dk/

[20] Gritton, B. et al., *Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government* (Working Papers). Report of the Ocean Sciences Data Panel; Commission on Physical Sciences, Mathematics, and Applications; National Research Council, National Academy Press, 1995, pp. 86-104.

[21] NOAA Administrative Order 216-101, *Ocean Data Acquisitions*, NOAA Office of Finance and Administration, July 9, 1990.

[22] Executive Order 12906, *Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure.* Federal Register, Vol. 59, No. 71, April 13, 1994, pp. 17671-17674.

[23] Based on email correspondence with Mr. Howard Diamond, Geospatial and Climate Services Group Leader, NOAA/NESDIS Office of the Chief Information Officer, January 22, 2002.

[24] http://badger.state.wi.us/agencies/wlib/sco/metatool/mtools.htm, http://www.fgdc.gov/metadata/toollist, http://umesc.usgs.gov/metamaker/nbiimker.html, http://www.intergraph.com/gis/smms, http://edcnts11.cr.usgs.gov/metalite, http://geology.usgs.gov/tools/metadata/tools/doc/tkme.html, http://www.blueangeltech.com

[25] http://fgdc.er.usgs.gov/

[26] Brabb, G. J., *Computers and Information Systems in Business*. Boston, Houghton Mifflin Co., 1987.

[27] Dozier et al., *Preserving Scientific Data on Our Physical Universe*. National Research Council, National Academy of Sciences Press, 1995.

[28] *Ibid*.

[29] *Op. Cit., Data Management for Marine Geology and Geophysics.*

[30] *Op. Cit.,* Wright, D.J. et al.

[31] Gritton, B.R. and C.H. Baxter, *Video Database Systems in the Marine Sciences*. Marine Technology Society Journal, 26(4), 1992, pp. 59-72.

[32] *Op. Cit.,* Wright, D.J. et al., p.2.

[33] *Principles for Dissemination of Scientific Data.* Report by the ad-hoc group on Data and Information, International Council for Science (ICSU) and the Committee on Data for Science and Technology (CODATA) at the World Intellectual Property Organization (WIPO) Information Meeting on Database Protection, Geneva, Switzerland, September 17-19, 1997.

[34] Current federal policies have been the subject of recent challenges resulting from a developing tension between full and open access policies and the protectionist desires of many commercial and nationalistic interests. Public Law 105-277 (the Omnibus Appropriations Act of 1998) directed that OMB apply Freedom of Information Act (FOIA) procedures to data produced under federal awards for the purpose of improving dissemination of data collected with federal support. The research community quickly realized that individuals might use the FOIA to compel access to raw, partially processed data that could disrupt the scientific process of discovery. As a result of intensive lobbying, the Circular A-110 was revised to clarify the statute as not requiring scientists to make federally funded research data publicly available while the research is still ongoing. The OMB goal in this revision was a balanced approach that would further the interest of the public in obtaining information needed to validate federally funded research findings, ensure that research could be conducted using the traditional scientific process, and implement a public access process that would be workable in practice. In parallel with this OMB action, both houses of Congress and the administration have considered new database protection measures championed by the publishing industry and international information conglomerates. These measures could lead to a more restricted environment for data collection, exchange, and use. In particular, enactment of U.S. database legislation could reduce the amount of data available to the public from the private sector or public-private partnerships and increase restrictions on the use of compilations of all kinds, including works of authorship (e.g., collections of articles) not normally considered to be databases. Further restrictions on the acquisition and use of data could be placed on researchers by risk-adverse universities and government agencies that would likely seek to avoid the possibility of costly litigation. The possible result could be a legal culture that encourages commercial exploitation rather than open access to information in the public domain.

[35] While NOAA Administrative Order 216-101 requires submission to national data management centers within one year of collection, it also provides for longer periods of time to accommodate particular data sets. These specific applications were to be identified by a Standing Committee for Ocean Data Policy; however, there is no record of such a committee publishing guidance on exceptions to the one-year policy. Within the federal government, general academic courtesy standards relating to federally funded research data have typically been manifested in policies similar to that of the National Science Foundation, which requires submission of a subset of metadata within 60 days of data collection and submission of data to the national data management centers within two years of collection.

[36] The Joint HDSA/NIST Data Preservation Test Facility is a cooperative effort by these two organizations. The HDSA charter is to understand and promote the use of automated, high-density data storage for backup and near-line recording and retrieval while the NIST charter addresses issues regarding the preservation of information and provision of access to this information. The facility is used to address interoperability, connectivity, and backward compatibility issues as well as to characterize performance and other specifications of high-density automated storage hardware, media, software, and related

systems. The facility also serves to stimulate economic growth of such data storage and retrieval systems through the use of hands-on demonstrations and published documents.

[37] Extracted from Draft Cruise Instructions for the Islands in the Stream Expedition, South Atlantic Bight Mission, Harbor Branch Oceanographic Institution and NOAA National Marine Sanctuaries, July 30, 2001.

[38] *Op. Cit.,* Stinus, et al.

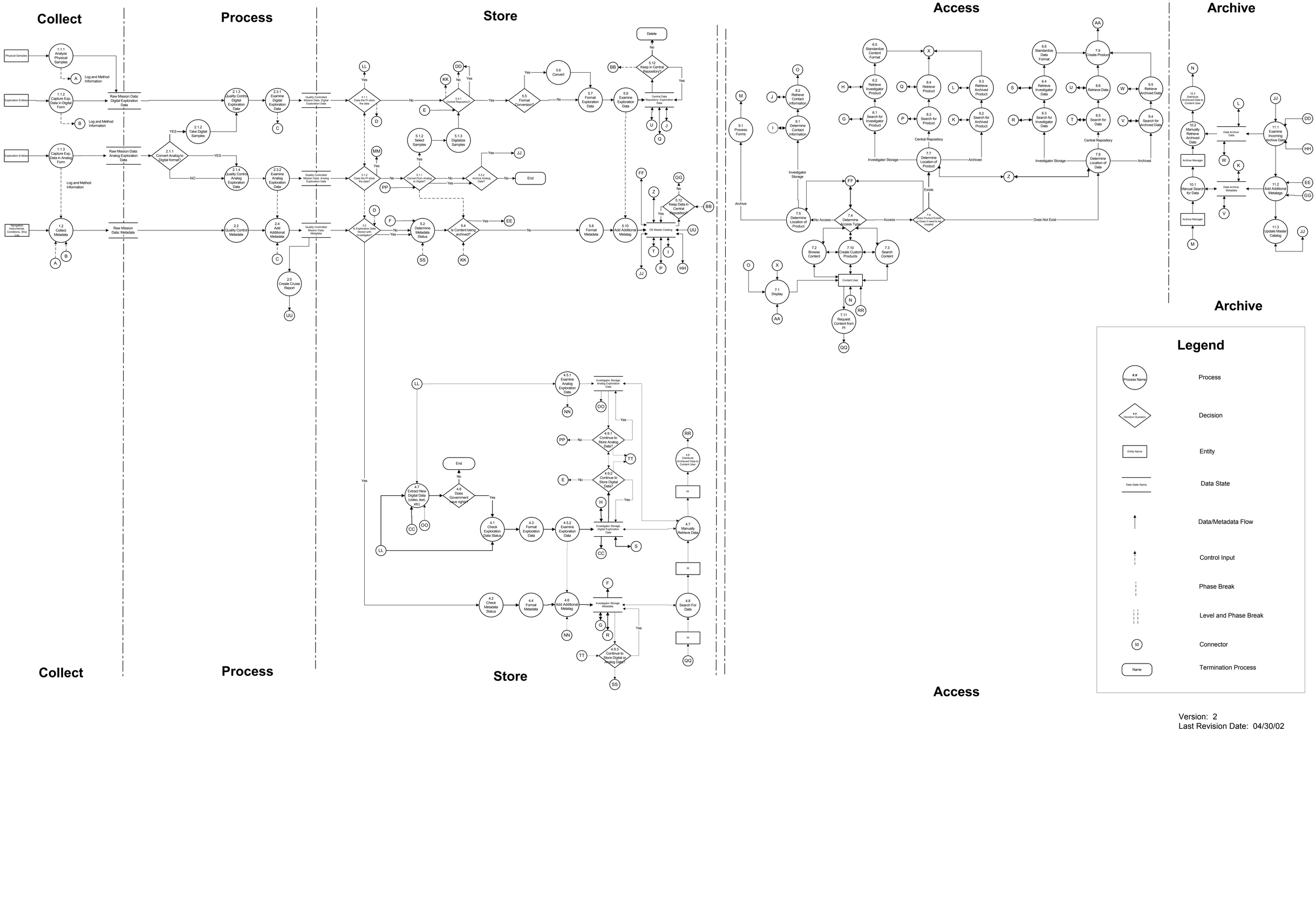
[39] *Op. Cit., Data Management for Marine Geology and Geophysics.*

[40] Cornillon, P. et al., *Possibilities of an Interoperable Data System: The DODS Model*. Presentation at the Oceanology International Americas Symposium, April 4, 2001; http://www.unidata.ucar.edu/packages/dods/

[42] Shea, D.J. et al., *An Introduction to Atmospheric and Oceanographic Data.* NCAR Technical Note TN-404-1A, National Centers for Atmospheric Research, Boulder, CO, August 1994.

[43] http://www.seg.org/publications/tech-stand/index_body.html

# Figure 3-2. OE Data Flow Model




Version: 2
Last Revision Date: 04/30/02


EXPLORE
NOAA